

UNIVERSAL
LIBRARY

OU_164189

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 510.6 / H51A Accession No. 38863

Author Henkin, L. and others.

Title Axiomatic method 1959

This book should be returned on or before the date
last marked below.

THE AXIOMATIC METHOD

STUDIES IN LOGIC AND THE FOUNDATIONS OF MATHEMATICS

L. E. J. BROUWER
E. W. BETH
A. HEYTING

Editors



1959

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM

THE AXIOMATIC METHOD

WITH SPECIAL REFERENCE TO GEOMETRY
AND PHYSICS

Proceedings of an International Symposium held at the
University of California, Berkeley, December 26, 1957 — January 4, 1958

Edited by

LEON HENKIN

Professor of Mathematics, University of California, Berkeley

PATRICK SUPPES

Associate Professor of Philosophy, Stanford University

ALFRED TARSKI

*Professor of Mathematics and Research Professor, University
of California, Berkeley*



1959

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM

*No part of this book may be reproduced
in any form by print, microfilm or any
other means without written permission
from the publisher*

PRINTED IN THE NETHERLANDS

CONTENTS

PREFACE	VII
-------------------	-----

PART I. FOUNDATIONS OF GEOMETRY

DIE MANNIGFALTIGKEIT DER DIREKTIVEN FÜR DIE GESTALTUNG GEOMETRISCHER AXIOMENSYSTEME. Paul Bernays	1
WHAT IS ELEMENTARY GEOMETRY? Alfred Tarski	16
SOME METAMATHEMATICAL PROBLEMS CONCERNING ELEMENTARY HYPERBOLIC GEOMETRY. Wanda Szmielew	30
DIMENSION IN ELEMENTARY EUCLIDEAN GEOMETRY. Dana Scott	53
BINARY RELATIONS AS PRIMITIVE NOTIONS IN ELEMENTARY GEOMETRY. Raphael M. Robinson	68
REMARKS ON PRIMITIVE NOTIONS FOR ELEMENTARY EUCLIDEAN AND NON-EUCLIDEAN PLANE GEOMETRY. H. L. Royden	86
DIRECT INTRODUCTION OF WEIERSTRASS HOMOGENEOUS COORDINATES IN THE HYPERBOLIC PLANE, ON THE BASIS OF THE ENDCALCULUS OF HILBERT. Paul Szász	97
AXIOMATISCHER AUFBAU DER EBENEN ABSOLUTEN GEOMETRIE. Friedrich Bachmann	114
NEW METRIC POSTULATES FOR ELLIPTIC n -SPACE. Leonard M. Blumenthal	127
AXIOMS FOR GEODESICS AND THEIR IMPLICATIONS. Herbert Busemann	146
AXIOMS FOR INTUITIONISTIC PLANE AFFINE GEOMETRY. A. Heyting	160
GRUNDLAGEN DER GEOMETRIE VOM STANDPUNKT DER ALLGEMEINEN TOPOLOGIE AUS. Karol Borsuk	174
LATTICE-THEORETIC APPROACH TO PROJECTIVE AND AFFINE GEOMETRY. Bjarni Jónsson	188
CONVENTIONALISM IN GEOMETRY. Adolf Grünbaum	204

PART II. FOUNDATIONS OF PHYSICS

HOW MUCH RIGOR IS POSSIBLE IN PHYSICS? Percy W. Bridgman	225
LA FINITUDE EN MÉCANIQUE CLASSIQUE, SES AXIOMES ET LEURS IMPLICATIONS. Alexandre Froda	238
THE FOUNDATIONS OF RIGID BODY MECHANICS AND THE DERIVATION OF ITS LAWS FROM THOSE OF PARTICLE MECHANICS. Ernest W. Adams	250
THE FOUNDATIONS OF CLASSICAL MECHANICS IN THE LIGHT OF RECENT AD- VANCES IN CONTINUUM MECHANICS. Walter Noll	266
ZUR AXIOMATISIERUNG DER MECHANIK. Hans Hermes	282
AXIOMS FOR RELATIVISTIC KINEMATICS WITH OR WITHOUT PARITY. Patrick Suppes	291
AXIOMS FOR COSMOLOGY. A. G. Walker	308
AXIOMATIC METHOD AND THEORY OF RELATIVITY. EQUIVALENT OBSERVERS AND SPECIAL PRINCIPLE OF RELATIVITY. Yoshio Ueno	322
ON THE FOUNDATIONS OF QUANTUM MECHANICS. Herman Rubin	333
THE MATHEMATICAL MEANING OF OPERATIONALISM IN QUANTUM MECHANICS. I. E. Segal	341
QUANTUM THEORY FROM NON-QUANTAL POSTULATES. Alfred Landé	353
QUANTENLOGIK UND DAS KOMMUTATIVE GESETZ. Pascual Jordan	365
LOGICAL STRUCTURE OF PHYSICAL THEORIES. Paulette Février	376
PHYSICO-LOGICAL PROBLEMS. J. L. Destouches	390

PART III. GENERAL PROBLEMS AND APPLICATIONS OF THE AXIOMATIC METHOD

STUDIES IN THE FOUNDATIONS OF GENETICS. J. H. Woodger	408
AXIOMATIZING A SCIENTIFIC SYSTEM BY AXIOMS IN THE FORM OF IDENTIFI- CATIONS. R. B. Braithwaite	429
DEFINABLE TERMS AND PRIMITIVES IN AXIOM SYSTEMS. Herbert A. Simon	443
AN AXIOMATIC THEORY OF FUNCTIONS AND FLUENTS. Karl Menger	454
AXIOMATICS AND THE DEVELOPMENT OF CREATIVE TALENT. R. L. Wilder	474

PREFACE

The thirty-three papers in this volume constitute the proceedings of an international symposium on *The axiomatic method, with special reference to geometry and physics*. This symposium was held on the Berkeley campus of the University of California during the period from December 26, 1957 to January 4, 1958.

The volume naturally divides into three parts. Part I consists of fourteen papers on the foundations of geometry, Part II of fourteen papers on the foundations of physics, and Part III of five papers on general problems and applications of the axiomatic method. General differences between the character of the papers in Part I and those in Part II reflect the relative state of development of the axiomatic method in geometry and in physics. Indeed, one of the important aims of the symposium was precisely to confront two disciplines in which the pattern of axiomatic development has been so markedly different.

Geometry, as is well known, is the science in which, more than 2300 years ago, the axiomatic method originated. Work on the axiomatization of geometry was greatly stimulated, and our conception of the significance and scope of the axiomatic method itself was greatly expanded, through the construction of non-Euclidean geometries in the first half of the nineteenth century. By the turn of the century we find for the first time, in the works of men like Pasch, Peano, Pieri, and Hilbert, axiomatic treatments of geometry which both are complete and meet the exacting standards of contemporary methodology of the deductive sciences. Since that time there has been a continuous and accelerating development of the subject, so that at present, all of the important geometric theories have been axiomatized, and new theories have been created through changes introduced into various systems of axioms; for most theories a variety of axiom systems is available conforming to varying ideals which have been pursued in connection with the axiomatization of geometry. Most recently, building upon the work on axiomatization, it has become possible to formalize geometrical theories, and in consequence geometrical theories themselves have been made the object of exact investigation by metamathematical methods, leading to several new kinds of results. The present volume contains new contri-

butions to many of the directions in which studies in the foundations of geometry have been developing.

Axiomatic work in the foundation of physics has had a more checkered history. Newton's *Principia*, first published in 1687, emulated Euclid's *Elements*, but the eighteenth and nineteenth centuries did not witness a development of axiomatic methods in physics at all comparable to that in geometry. Even the work in this century on axiomatizing various branches of physics has been relatively slight in comparison with the massive mathematical development of geometry. There is not to our knowledge a single treatise on classical mechanics which compares in axiomatic precision with such a work as Hilbert's well-known text on the foundations of geometry; furthermore, the axiomatic treatments of various branches of physics which have been attempted, including those presented in this volume, do not yet have the finished and complete character typical of geometrical axiomatizations. Much foundational work in physics is still of the programmatic sort, and it is possible to maintain that the status of axiomatic investigations in physics is not yet past the preliminary stage of philosophical discussion expressing doubt as to its purpose and usefulness. In spite of such doubts, an increasing effort is being made to apply axiomatic methods in physics, and many of the papers in Part II indicate how exact mathematical methods may be brought to bear on problems in the foundations of physics. To the knowledge of the Editors the papers in Part II constitute the first collection whose aim is specifically to provide an over-all perspective of the application of axiomatic methods in physics. It is our candid hope that this book will be a stimulus to further work in this important domain.

An attempt has been made to give coherence to the volume by grouping papers according to their subject. Part I begins with a paper by Bernays which surveys the main tendencies manifesting themselves in the construction of geometrical axiom systems. In the five papers which follow, metamathematical notions referring to the axiomatic foundations of various systems of Euclidean and non-Euclidean geometry are discussed, and to a large extent specific metamathematical methods are applied. The first three papers, namely those by Tarski, Szmielew, and Scott, are concerned with problems of completeness and decidability, while the remaining two, those by Robinson and Royden, deal with problems of definability. The next five papers set forth new axiomatizations of various branches of geometry. Szász is concerned with hyperbolic geometry, Bachmann with absolute geometry, Blumenthal with elliptic geometry,

Busemann with metric differential geometry, and Heyting with affine geometry; the last of these authors approaches the subject from the intuitionistic point of view. In the following two papers connections between the foundations of geometry and some related branches of mathematics are studied. In particular, Borsuk examines Euclidean geometry from the standpoint of topology, and Jónsson surveys projective and affine geometry from the standpoint of lattice theory. The last paper of Part I, that of Grünbaum, deals with the philosophical problem of conventionalism in geometry.

In the case of Part II, the order has, roughly speaking, followed the historical development of physics. The opening paper of Bridgman analyzes the general notion of rigor in physics. It is followed by four papers on the axiomatic foundations of classical mechanics. Froda considers particle mechanics, Adams rigid body mechanics and Noll continuum mechanics; Hermes analyzes certain axiomatic problems surrounding the notion of mass. Three papers on relativity follow. Suppes deals with relativistic kinematics, Walker with relativistic cosmology and Ueno with relativity theory as based on the concept of equivalent observers. Next come three papers on quantum mechanics. Rubin considers quantum mechanics from the standpoint of the theory of stochastic processes; Segal examines the mathematical meaning of operationalism in quantum mechanics; and Landé approaches the subject on the basis of non-quantal postulates. Finally, there are three papers which deal with relations between logic and physics. Jordan considers quantum logic and the commutative law; Février the logical structure of physical theories; and Destouches the theory of prediction with special reference to physico-logical problems.

The arrangement of papers in Part III is somewhat arbitrary. Loosely speaking, the papers move from more specific to more general topics. Woodger is concerned with the foundations of genetics; Braithwaite with scientific theories whose axioms take the form of identities; Simon with primitive and definable terms in axiom systems; Menger with the general theory of functions in the context of the empirical sciences; and Wilder with the potentiality of the axiomatic approach as a method of teaching.

It goes without saying that each author is solely responsible for the content of his paper. The Editors have confined themselves to arranging the volume and handling various technical matters relating to publication. In particular, the choice of notation and symbolism has been left to the individual author.

The calendar of the scientific sessions was as follows:

December 26, afternoon. Opening remarks by Acting Chancellor James D. Hart of the University of California, Berkeley, and by Professor Alfred Tarski of the same University. Section I, Professor Paul Bernays (Zürich, Switzerland). Section II, Professor P. W. Bridgman (Cambridge, Massachusetts, U.S.A.).

December 27, morning. Section II, Professor Hans Hermes (Münster, Germany), Professor Walter Noll (Pittsburgh, Pennsylvania, U.S.A.).

December 27, afternoon. Section I, Professor Friedrich Bachmann (Kiel, Germany), Professor Alfred Tarski (Berkeley, California, U.S.A.).

December 28, morning. Section II, Professor Ernest Adams (Berkeley, California, U.S.A.), Professor Yoshio Ueno (Hiroshima, Japan).

December 28, afternoon. Section I, Mr. Dana Scott (Princeton, New Jersey, U.S.A.), Professor H. L. Royden (Stanford, California, U.S.A.), Professor Raphael M. Robinson (Berkeley, California, U.S.A.).

December 30, morning. Section II, Professor Arthur G. Walker (Liverpool, England), Professor Patrick Suppes (Stanford, California, U.S.A.).

December 30, afternoon. Commemorative talks on the first anniversary of the death of Heinrich Scholz by Alfred Tarski and Paul Bernays. Section I, Professor Paul Szász (Budapest, Hungary), Professor Wanda Szmielew (Warsaw, Poland, and Berkeley, California, U. S.A.).

December 31, morning. Section II, Professor Irving E. Segal (Chicago, Illinois, U.S.A.), Professor Jean-Louis Destouches (Paris, France).

December 31, afternoon. Section I, Professor Leonard M. Blumenthal (Columbia, Missouri, U.S.A.), Professor Herbert Busemann (Los Angeles, California, U.S.A.).

January 2, morning. Section III, Professor Joseph H. Woodger (London, England), Professor Richard Braithwaite (Cambridge, England).

January 2, afternoon. Section I, Professor Karol Borsuk (Warsaw, Poland), Professor Bjarni Jónsson (Minneapolis, Minnesota, U.S.A.).

January 3, morning. Section II, Professor Pascual Jordan (Hamburg, Germany), Dr. Paulette Février (Paris, France).

January 3, afternoon. Section I, Professor Arend Heyting (Amsterdam, Netherlands), Professor Adolf Grünbaum (Bethlehem, Pennsylvania, U.S.A.).

January 4, morning. Section II, Professor Alfred Landé (Columbus, Ohio, U.S.A.), Professor Herman Rubin (Eugene, Oregon, U.S.A.).

January 4, afternoon. Section III, Professor Karl Menger (Chicago,

Illinois, U.S.A.), Professor Raymond L. Wilder (Ann Arbor, Michigan, U.S.A.).

Three invited speakers whose papers are included in this volume were unable actually to attend the symposium: the paper of Paul Szász was read by Steven Orey, and the papers of Alexandre Froda and Herbert Simon were presented by title. Several talks were presented originally under different titles than appear in this volume: R. B. Braithwaite, *Necessity and contingency in the empirical interpretation of axiomatic systems*; A. Landé, *Non-quantal foundations of quantum mechanics*; H. L. Royden, *Binary relations as primitive notions in geometry with set-theoretical basis*; A. G. Walker, *Axioms of kinematical relativity*.

This symposium was jointly sponsored by the U. S. National Science Foundation (which contributed the bulk of the supporting funds), the International Union for the History and Philosophy of Science (Division of Logic, Methodology, and Philosophy of Science), and the University of California. The symposium was organized by a committee consisting of Leon Henkin, Secretary (University of California, Berkeley), Victor F. Lenzen (University of California, Berkeley), Benson Mates (University of California, Berkeley), Ernest Nagel (Columbia University, New York), Steven Orey (University of California, Berkeley, and University of Minnesota, Minneapolis), Julia Robinson (Berkeley, California), Patrick Suppes (Stanford University, Stanford, California), Alfred Tarski, Chairman (University of California, Berkeley), and Raymond L. Wilder (University of Michigan, Ann Arbor). The Secretary of the symposium was Dorothy Wolfe.

We gratefully acknowledge the help of Mr. Rudolf Grewe and Dr. Dana Scott in preparing this volume for publication.

University of California, Berkeley
Stanford University
February 1959

THE EDITORS

DIE MANNIGFALTIGKEIT DER DIREKTIVEN FÜR DIE GESTALTUNG GEOMETRISCHER AXIOMENSYSTEME

PAUL BERNAYS

Eidgenössische Technische Hochschule, Zürich, Schweiz

Bei der Betrachtung der Axiomatisierungen der Geometrie stehen wir unter dem Eindruck der grossen Mannigfaltigkeit der Gesichtspunkte, unter denen die Axiomatisierung erfolgen kann und auch schon erfolgte. Die ursprüngliche einfache alte Vorstellung, wonach man schlechtweg von *den* Axiomen der Geometrie sprechen kann, ist nicht nur durch die Entdeckung der nichteuklidischen Geometrien verdrängt, und ferner auch durch die Einsicht in die Möglichkeit verschiedener Axiomatisierungen einer und derselben Geometrie, sondern es sind überhaupt wesentlich verschiedene methodische Gesichtspunkte aufgetreten, unter denen man die Axiomatisierung der Geometrie unternommen hat und deren Zielsetzungen sogar in gewissen Beziehungen antagonistisch sind.

Der Keim für diese Mannigfaltigkeit ist bereits in der euklidischen Axiomatik zu finden. Für deren Gestaltung war der Umstand bestimmend, dass man hier an Hand der Geometrie zum ersten Mal auf die Problemstellung der Axiomatik geführt wurde. Die Geometrie ist hier sozusagen die Mathematik schlechthin. Das Verhältnis zur Zahlentheorie ist methodisch wohl kein völlig deutliches. In gewissen Teilen wird ein Stück Zahlentheorie mit Verwendung der anschaulichen Zahlvorstellung entwickelt. Ferner wird in der Proportionenlehre inhaltlich von dem Zahlbegriff Gebrauch gemacht, sogar mit einem impliziten Einschluss des Tertium non datur; allerdings scheint es, dass man dessen volle Verwendung zu vermeiden trachtete.

Während die methodische Sonderstellung des Zahlbegriffes hier nicht explizite hervortritt, wird der Grössenbegriff ausdrücklich als inhaltliches Hilfsmittel an die Spitze gestellt, in einer Art übrigens, die wir heute nicht mehr konzedieren können, indem nämlich von verschiedenen Gegenständlichkeiten als selbstverständlich vorausgesetzt wird, dass sie Grössencharakter haben. Der Grössenbegriff wird freilich auch der Axiomatisierung unterworfen; die diesbezüglichen Axiome werden jedoch ausdrücklich als vorgängige (*kouai êrvoiai*) von den übrigen Axiomen abgesondert.

Diese Axiome sind von ähnlicher Art, wie diejenigen, die man heute für die abelschen Gruppen aufstellt. Was aber auf Grund des damaligen methodischen Standpunktes unterblieb, war, dass nicht axiomatisch fixiert wurde, welche Gegenstände als Grössen anzusehen seien.

Umsomehr ist es zu bewundern, dass man damals schon auf das Besondere derjenigen Voraussetzung aufmerksam wurde, durch welche die archimedischen Grössen, wie wir sie heute nennen, ausgezeichnet werden. Das Archimedische (Eudoxische) Axiom wird dann, in der an die Griechen anschliessenden mittelalterlichen Tradition, insbesondere in den Untersuchungen der Araber über das Parallelenaxiom wesentlich benutzt. Auch bei dem Beweis von Saccheri zur Ausschliessung der „Hypothese des stumpfen Winkels“ tritt es als wesentlich auf. In der Tat ist ja diese Ausschliessung ohne das Archimedische Axiom nicht möglich, da ja eine nicht-archimedische, schwach-sphärische (bzw. schwach-elliptische) Geometrie mit den Axiomen der euklidischen Geometrie, abgesehen vom Parallelenaxiom, im Einklang steht.

Bei allen diesen Untersuchungen tritt das zweite Stetigkeitsaxiom, welches im späteren 19.ten Jahrhundert formuliert wurde, noch nicht auf. Es konnte bei den Beweisführungen, für die es in Betracht kam — wie bei den Flächeninhalts- und Längenbestimmungen — auf Grund der erwähnten Verwendung des Grössenbegriffs, entbehrt werden, wonach es z.B. als selbstverständlich galt, dass die Kreisfläche sowie der Kreisumfang eine bestimmte Grösse besitzen. An die Stelle der alten Grössenlehre trat zum Beginn der Neuzeit als beherrschende übergeordnete Disziplin die Grössenlehre der *Analysis*, die sich formal und dem Inhalt nach sehr reich entwickelte, noch ehe sie zu methodischer Deutlichkeit gelangte.

Freilich, bei der Entdeckung der nichteuklidischen Geometrie spielte die Analysis zunächst keine erhebliche Rolle, wohl aber wird sie dominierend in den nachfolgenden Untersuchungen von Riemann und Helmholtz, und später von Lie, zur Kennzeichnung der drei ausgezeichneten Geometrien durch gewisse sehr allgemeine, analytisch fassbare Bedingungen. Charakteristisch für diese Behandlung der Geometrie ist insbesondere, dass man nicht nur die einzelnen Raumgebilde, sondern auch die Raummannigfaltigkeit selbst zum Gegenstand nimmt. In der Möglichkeit der Durchführung einer solchen Betrachtung zeigten sich die gewaltigen begrifflichen und formalen Mittel, welche die Mathematik in der Zwischenzeit gewonnen hatte; und in der Anlage der Problemstellung äusserte sich die begrifflich-spekulative Richtung, welche die Mathematik im Laufe des 19.ten Jahrhunderts einschlug.

Die differentialgeometrische Behandlung der Grundlagen der Geometrie ist ja übrigens bis in die neueste Zeit durch Hermann Weyl sowie Elie Cartan und Levi-Civita, in Anknüpfung an die allgemeine Relativitätstheorie Einsteins, weiter entwickelt worden. So imponierend und elegant das in dieser Hinsicht Erreichte ist, so haben sich doch die Mathematiker vom grundlagentheoretischen Standpunkt damit nicht zufrieden gegeben. Zunächst suchte man sich von der für die differentialgeometrische Methode wesentlichen Voraussetzung der Differenzierbarkeit der Abbildungsfunktionen zu befreien. Dafür bedurfte es der Ausbildung der Methoden einer allgemeinen Topologie, welche um die Wende des Jahrhunderts begann und seitdem eine so imposante Entwicklung genommen hat. Weitergehend trachtete man sich von der Voraussetzung des archimedischen Charakters der geometrischen Grössen überhaupt unabhängig zu machen.

Diese Tendenz steht im Zeichen derjenigen Entwicklung, mit welcher die Analysis ihre vorher beherrschende Stellung in gewissem Masse eingebüsst hat. Dieses neue Stadium in der mathematischen Forschung knüpfte sich an die Auswirkung der schon erwähnten begrifflich-spekulativen Richtung der Mathematik im 19.ten Jahrhundert, wie sie insbesondere in der Schöpfung der allgemeinen Mengenlehre, in der schärferen Begründung der Analysis, in der Konstitution der mathematischen Logik und in der neuen Fassung der Axiomatik in Erscheinung trat.

Für dieses neue Stadium war zugleich charakteristisch, dass man wieder mehr auf die Methoden der alten griechischen Axiomatik zurückkam, wie es wiederholt in den Epochen geschah, in denen man auf begriffliche Präzision stärkeren Nachdruck legte. In Hilberts Grundlagen der Geometrie finden wir einerseits dieses Zurückkommen auf die alte elementare Axiomatik, freilich in grundsätzlich veränderter methodischer Auffassung, andererseits als ein hauptsächliches Thema die möglichst weitgehende Ausschaltung des archimedischen Axioms: sowohl bei der Proportionenlehre wie beim Flächeninhaltsbegriff sowie in der Begründung der Streckenrechnung. Diese Art der Axiomatisierung hatte übrigens für Hilbert nicht den Sinn der Ausschliesslichkeit; er hat ja bald danach eine andere Art der Begründung daneben gestellt, mit der zum ersten Mal das vorhin erwähnte Programm einer topologischen Grundlegung aufgestellt und durchgeführt wurde.

Etwa gleichzeitig mit Hilberts Grundlegung wurde auch in der Schule von Peano und Pieri die Axiomatisierung der Geometrie gepflegt. Bald folgten auch die axiomatischen Untersuchungen von Veblen und R. L.

Moore; und es waren nunmehr die Forschungsrichtungen eingeschlagen, in denen sich auch heute die Beschäftigung mit den Grundlagen der Geometrie weiterbewegt. Als kennzeichnend hierfür haben wir eine Vielheit der methodischen Richtungen.

Die eine ist die, welche die Mannigfaltigkeit der kongruenten Transformationen durch möglichst allgemeine und prägnante Bedingungen zu kennzeichnen sucht, die zweite, diejenige, welche die projektive Struktur des Raumes voranstellt und das Metrische auf das Projektive mit der von Cayley und Klein ausgebildeten Methode der projektiven Massbestimmung zurückzuführen trachtet, und die dritte die, welche auf eine elementare Axiomatisierung der vollen Kongruenzgeometrie ausgeht.

Verschiedene wesentlich neue Gesichtspunkte sind in der Entwicklung dieser Richtungen hinzugetreten. Einmal erhielt die projektive Axiomatik eine verstärkte Systematisierung mittels der Verbandstheorie. Ferner wurde man gewahr, dass man bei der Kennzeichnung der Gruppe der kongruenten Transformationen die mengentheoretischen und funktionentheoretischen Begriffsbildungen zurücktreten lassen kann, indem man die Transformationen durch sie bestimmende Gebilde festlegt. Damit kommt das Verfahren dem der elementaren Axiomatik nahe, da die Gruppenbeziehungen sich nun als Beziehungen zwischen geometrischen Gebilden darstellen.

Ich will aber hier nicht näher von diesen beiden Forschungsrichtungen der geometrischen Axiomatik sprechen, für die ja hier authentischere Vertreter anwesend sind, auch nicht von den Erfolgen, die mit Verwendung topologischer Methoden erzielt worden sind, worüber insbesondere neuere Abhandlungen von Freudenthal einen Überblick liefern, sondern mich den Fragen der an dritten Stelle genannten Richtung der Axiomatisierung zuwenden.

Selbst innerhalb dieser Richtung finden wir wiederum eine Mannigfaltigkeit von möglichen Zielsetzungen. Man kann einerseits darauf ausgehen, mit möglichst wenigen Grundelementen, etwa nur einem Grundprädikat und einer Gattung von Individuen, auszukommen. Andererseits kann man vornehmlich darauf gerichtet sein, natürliche Absonderungen von Teilen der Axiomatik hervortreten zu lassen. Diese Gesichtspunkte führen zu verschiedenen Alternativen.

So wird einerseits durch die Betrachtung der nichteuklidischen Geometrie die Voranstellung der „absoluten“ Geometrie nahegelegt. Andererseits hat auch ein solcher Aufbau manches für sich, bei dem die affine Vektorgeometrie vorangestellt wird, wie es am Anfang von Weyl's

„Raum, Zeit, Materie“ geschieht. Diesen beiden Gesichtspunkten kann man schwerlich zugleich in einer Axiomatik Genüge tun. Ein anderes Beispiel ist dieses. Bei der Voranstellung der Axiome der Inzidenz und Anordnung ist es eine mögliche und elegante begriffliche Reduktion, dass man, nach dem Vorgang von Veblen, den Begriff der Kollinearität auf den Zwischen-Begriff zurückführt. Andererseits ist es für manche Überlegungen von Wichtigkeit, die von dem Anordnungsbegriff unabhängigen Folgerungen der Inzidenzaxiome abzusondern; so ist es ja wünschenswert die Begründung der Streckenrechnung aus den Inzidenzaxiomen als unabhängig von den Anordnungsaxiomen zu erkennen. Wiederum bei der Theorie der Anordnung selbst hat man Ersparungen von Axiomen der linearen Anordnung durch Anwendung des Axioms von Pasch als möglich erkannt; andererseits ist in gewisser Hinsicht eine Anlage der Axiome zu bevorzugen, bei welcher die für die lineare Anordnung kennzeichnenden Axiome abgesondert werden.

Mit diesen Beispielen von Alternativen ist die Mannigfaltigkeit in den möglichen und auch den tatsächlich verfolgten Zielsetzungen nicht annähernd erschöpft. So ist es ein möglicher und sinngemässer, wenn auch nicht obligatorischer regulativer Gesichtspunkt, dass die Axiome so formuliert werden sollen, dass sie sich jeweils nur auf ein beschränktes Raumstück beziehen. Dieser Gedanke ist implicite ja wohl schon in der euklidischen Axiomatik mitbestimmend; und es mag auch sein, dass der Anstoss, den man so frühzeitig an dem Parallelenaxiom genommen hat, gerade darauf beruht, dass in der euklidischen Formulierung der Begriff der genügend weiten Verlängerung auftritt. Die erstmalige explizite Durchführung des genannten Programmpunktes geschah durch Moritz Pasch, und es knüpfte sich daran die Einführung idealer Elemente mit Hilfe von Schnittpunktsätzen, eine seitdem in erfolgreicher Weise ausgestaltete Methode der Begründung der projektiven Geometrie.

Eine andere Art der möglichen zusätzlichen Aufgabestellung ist diejenige, die Unschärfe unseres bildhaften Vorstellens begrifflich nachzuahmen, wie dieses ja Hjelmslev getan hat. Das ergibt freilich nicht nur eine andere Art der Axiomatisierung, sondern überhaupt ein abweichendes Beziehungssystem, ein Verfahren, welches wohl wegen seiner Komplikation nicht viel Anklang gefunden hat. Doch auch ohne in dieser Richtung sich soweit von dem Üblichen zu entfernen, kann man etwas in gewisser Hinsicht Ähnliches anstreben, indem man den Begriff des Punktes als Gattungsbegriff vermeidet, wie es ja in verschiedenen interessanten

nuereen Axiomatisierungen geschieht, so insbesondere in derjenigen von Huntington.

In solcher Weise zeigt sich auf mannigfachste Art, dass es kein eindeutiges Optimum für die Gestaltung eines geometrischen Axiomensystems gibt. Was übrigens die Reduktionen in Hinsicht der Grundbegriffe und der Dingarten betrifft, so ist ungeachtet des grundsätzlichen Interesses, welches jede solche Reduktionsmöglichkeit hat, doch immer daran zu erinnern, dass die tatsächliche Anwendung einer solchen Reduktion sich nur dann empfiehlt, wenn damit eine übersichtliche Gestaltung des Axiomensystems erreicht wird.

Es lassen sich immerhin gewisse Direktiven für Reduktionen nennen, die wir generell akzeptieren können. Nehmen wir etwa als Beispiel die Hilbert'sche Fassung der Axiomatik. Bei dieser werden einerseits die Geraden als eine Dinggattung genommen, andererseits die Halbstrahlen als Punktmengen eingeführt und anschliessend dann die Winkel als geordnete Paare zweier von einem Punkt ausgehender Halbstrahlen, also als Paare von Mengen, erklärt. Hier sind tatsächlich Möglichkeiten der vereinfachenden Reduktion gegeben. Man mag verschiedener Meinung darüber sein, ob man anstatt der verschiedenen Gattungen „Punkt, Gerade, Ebene“ nur eine Gattung der Punkte zugrunde legen will, wobei dann anstelle der Inzidenzbeziehung die Beziehungen der Kollinearität und der Komplanarität von Punkten treten. In der verbandstheoretischen Behandlung werden ja die Geraden und Ebenen gleichstehend mit den Punkten als Dinge genommen. Hier steht man wiederum vor einer Alternative. Hingegen die Halbstrahlen als Punktmengen einzuführen, überschreitet jedenfalls den Rahmen der elementaren Geometrie und ist auch für diese nicht nötig. Generell können wir es wohl als Direktive nehmen, dass höhere Gattungen nicht ohne Erfordernis eingeführt werden sollen. Beim Fall der Winkeldefinition kann man das dadurch vermeiden, dass man die Winkelaussagen auf Aussagen über Punkttupel reduziert, wie dieses ja von R. L. Moore durchgeführt wurde. Hier wird sogar noch eine weitere Reduktion erreicht, indem überhaupt die Winkelkongruenz mit Hilfe der Streckenkongruenz erklärt wird, doch findet hierbei wiederum auch eine gewisse Einbusse statt. Nämlich die Beweisführungen stützen sich dabei wesentlich auf die Kongruenz von ungleichsinnig zugeordneten Dreiecken. Daher ist diese Art der Axiomatisierung nicht geeignet für den Problemkreis derjenigen Hilbert'schen Untersuchungen, welche sich auf das Verhältnis der gleichsinnigen Kongruenz zur Symmetrie beziehen. Diese Bemerkung betrifft freilich auch die meisten der

Axiomatisierungen, bei denen der Begriff der Spiegelungen an der Spitze steht.

Neben den allgemeinen Gesichtspunkten möchte ich als etwas Einzelnes eine spezielle Möglichkeit der Anlage eines elementaren Axiomensystems erwähnen, nämlich eine solche Axiomatik, bei welcher der Begriff „das Punktetripel a, b, c bildet bei b einen rechten Winkel“ als einzige Grundbeziehung und die Punkte als einzige Grundgattung genommen werden, ein Programm, auf welches neuerdings durch eine Arbeit von Dana Scott hingewiesen worden ist. Die genannte Beziehung genügt der von Tarski festgestellten notwendigen Bedingung für ein allein ausreichendes Grundprädikat der Planimetrie. Im Vergleich mit dem für eine Axiomatik solcher Art vorbildlich gewordenen Verfahren Pieri's, der ja in einer Axiomatisierung die Beziehung „ b und c haben von a gleichen Abstand“ als Grundbegriff nahm, scheint hier insofern eine Erleichterung zu bestehen, als der Begriff der Kollinearität von Punkten sich enger an den des rechten Winkels als an den Pieri'schen Grundbegriff anschliesst. Was freilich den Kongruenzbegriff anbelangt, so scheint sich für die Axiome der Kongruenz aus der betrachteten Reduktion keine Vereinfachung zu ergeben. Übrigens ist diese Axiomatisierung ebenso wie die genannte Pieri'sche eine von denen, die keine Aussonderung der gleichsinnigen Kongruenz liefern ¹.

Für eine elementare Axiomatisierung der Geometrie stellt sich als besondere Frage die der Gewinnung einer Vollständigkeit im Sinne der Kategorizität. Diese wird bei den meisten Axiomensystemen durch die Stetigkeitsaxiome erwirkt. Die Einführung dieser Axiome bedeutet aber, wie man weiss, eine Überschreitung des Rahmens der gewöhnlichen Prädikatenlogik, indem das archimedische Axiom den allgemeinen Zahlbegriff verwendet und das zweite Stetigkeitsaxiom den allgemeinen Prädikaten- oder Mengenbegriff. Wir haben seither aus den Untersuchungen Tarski's gelernt, dass wir eine Vollständigkeit, wenigstens im deduktiven Sinne, in einem elementaren Rahmen erreichen können, wobei das Bemerkenswerte ist, dass das Schnittaxiom in einer gewissen Formalisierung erhalten bleibt, während von dem Archimedischen Axiom abgesehen wird. Das Archimedische Axiom fällt ja insofern formal aus dem sonstigen Rahmen heraus, als es in logischer Formalisierung die Gestalt einer un-

¹) Einige Angaben über die Definitionen der Inzidenz-, Anordnungs- und Kongruenzbegriffe aus dem Begriff des rechten Winkels, sowie über einen Teil des Axiomensystems folgen in einem Anhang.

endlichen Alternative hat, während das Schnittaxiom auf Grund seiner Form der Allgemeinheit sich durch ein Axiomenschema darstellen und dadurch in seiner Anwendung dem jeweiligen formalen Rahmen anpassen lässt, — wobei dann für den elementaren Rahmen der Prädikatenlogik die Beweisbarkeit des Archimedischen Axioms aus dem Schnittaxiom verloren geht. Freilich hat eine solche Beschränkung auf einen prädikatenlogischen Rahmen zur Folge, dass verschiedene Überlegungen nur metatheoretisch ausgeführt werden können, wie z.B. der Beweis des Satzes, dass ein einfach geschlossenes Polygon die Ebene zerlegt, und ebenso die Betrachtung über Ergänzungsgleichheit und Zerlegungsgleichheit von Polygonen. Man steht hier wieder einmal vor einer Alternative, nämlich der, ob man den Gesichtspunkt der Elementarität des logischen Rahmens voranstellen will, oder sich hinsichtlich des logischen Rahmens nicht beschränkt, wobei ja übrigens noch verschiedene Abstufungen in Betracht kommen.

In Bezug auf die Anwendung einer Logik der zweiten Stufe sei hier nur daran erinnert, dass eine solche sich ja im Rahmen der axiomatischen Mengenlehre in solcher Weise präzisieren lässt, dass keine fühlbare Einschränkung der Beweismethoden erfolgt. Auch das Skolem'sche Paradoxon bereitet im Falle der Geometrie insofern keine eigentliche Verlegenheit, als man es dadurch ausschalten kann, dass man in den modeltheoretischen Betrachtungen den Mengenbegriff, der in einem der höheren Axiome auftritt, mit dem Mengenbegriff der Modelltheorie gleichsetzt.

Zum Schluss möchte ich hervorheben, dass der in meinen Ausführungen betonte Umstand, dass es in der Gestaltung der Axiomatik kein eindeutiges Optimum gibt, keineswegs bedeutet, dass die Erzeugnisse der geometrischen Axiomatik notwendig den Charakter des Unvollkommenen und Fragmentarischen tragen. Sie wissen, dass auf diesem Gebiete etliche Gestaltungen von grosser Vollkommenheit und Abrundung erreicht worden sind. Gerade die Vielheit der möglichen Zielrichtungen bewirkt, dass durch das Neuere das Frühere im allgemeinen nicht schlechtweg überholt wird, während andererseits auch jede erreichte Vollkommenheit immer noch Platz lässt für weitere Aufgaben.

ANHANG. *Bemerkungen zu der Aufgabe einer Axiomatisierung der euklidischen Planimetrie mit der einzigen Grundbeziehung $R(a, b, c)$: „das Punktetripel a, b, c , bildet bei b einen rechten Winkel“.* Die Axiomatisierung gelingt insoweit auf einfache Art, als nur die Beziehungen der Kollinearität und des Parallelismus betrachtet werden. Für die Theorie der Koll-

nenität genügen die folgenden Axiome:

$$A1 \quad \neg R(a, b, a)$$

$$A2 \quad R(a, b, c) \rightarrow R(c, b, a) \ \& \ \neg R(a, c, b)^2$$

$$A3 \quad R(a, b, c) \ \& \ R(a, b, d) \ \& \ R(e, b, c) \rightarrow R(e, b, d)$$

$$A4 \quad R(a, b, c) \ \& \ R(a, b, d) \ \& \ c \neq d \ \& \ R(e, c, b) \rightarrow R(e, c, d)$$

$$A5 \quad a \neq b \rightarrow (Ex)R(a, b, x).$$

Dazu tritt die Definition der Beziehung $\text{Koll}(a, b, c)$: „die Punkte a, b, c sind kollinear“:

$$\text{DEFINITION 1.} \quad \text{Koll}(a, b, c) \leftrightarrow (x)(R(x, a, b) \rightarrow R(x, a, c)) \vee a = c.$$

Es sind dann die folgenden Sätze beweisbar:

$$(1) \quad \text{Koll}(a, b, c) \leftrightarrow a = b \vee a = c \vee b = c \vee (Ex)(R(x, a, b) \ \& \ R(x, a, c))$$

$$(2) \quad \text{Koll}(a, b, c) \rightarrow \text{Koll}(a, c, b) \ \& \ \text{Koll}(b, a, c)$$

$$(3) \quad \text{Koll}(a, b, c) \ \& \ \text{Koll}(a, b, d) \ \& \ a \neq b \rightarrow \text{Koll}(b, c, d)$$

$$(4) \quad R(a, b, c) \ \& \ \text{Koll}(b, c, d) \ \& \ b \neq d \rightarrow R(a, b, d)$$

$$(5) \quad R(a, b, c) \rightarrow \neg \text{Koll}(a, b, c)$$

$$(6) \quad R(a, b, c) \ \& \ R(a, b, d) \rightarrow \text{Koll}(b, c, d)$$

$$(7) \quad R(a, b, c) \ \& \ R(a, b, d) \rightarrow \neg R(a, c, d).$$

$$\text{Zum Beweis: } \text{Koll}(c, d, b) \ \& \ c \neq b \rightarrow (R(a, c, d) \rightarrow R(a, c, b))$$

$$(8) \quad R(a, b, c) \ \& \ R(a, b, d) \ \& \ R(a, e, c) \ \& \ R(a, e, d) \rightarrow c = d \vee b = e.$$

$$\text{Zum Beweis: } \text{Koll}(b, c, d) \ \& \ \text{Koll}(e, c, d) \ \& \ c \neq d \rightarrow \text{Koll}(b, c, e)$$

$$\text{Koll}(b, c, e) \ \& \ b \neq e \ \& \ R(a, b, c) \rightarrow R(a, b, e)$$

$$\text{Koll}(e, c, b) \ \& \ b \neq e \ \& \ R(a, e, c) \rightarrow R(a, e, b)$$

$$R(a, b, e) \rightarrow \neg R(a, e, b).$$

Für die Theorie des Parallelismus nehmen wir zwei weitere Axiome hinzu:

$$A6 \quad a \neq b \ \& \ a \neq c \rightarrow (Ex)(R(x, a, b) \ \& \ R(x, a, c)) \vee$$

$$(Ex)(R(a, x, b) \ \& \ R(a, x, c)) \vee R(a, b, c) \vee R(a, c, b)$$

²⁾ Durch dieses Axiom wird bereits die elliptische Geometrie ausgeschlossen.

Das Axiom besagt in üblicher Ausdrucksweise, dass man von einem Punkte a ausserhalb einer Geraden bc auf diese eine Senkrechte fallen kann. Die eindeutige Bestimmtheit der Senkrechten in Abhängigkeit von dem Punkt a und der Geraden bc ergibt sich mit Hilfe von (4) und (8).

$$A7 \quad R(a, b, c) \& R(b, c, d) \& R(c, d, a) \rightarrow R(d, a, b)$$

Dieses ist eine Form des euklidischen Parallelenaxioms im engeren, winkelmetrischen Sinn.

Die Parallelität wird nun definiert durch:

$$\text{DEFINITION 2. } \text{Par}(a, b, c, d) \leftrightarrow a \neq b \& c \neq d \& (Ex)(Ey)(R(a, x, y) \& R(b, x, y) \& R(c, y, x) \& R(d, y, x))$$

Als beweisbare Sätze ergeben sich:

- (9) $\text{Par}(a, b; c, d) \rightarrow \text{Par}(b, a; c, d) \& \text{Par}(c, d; a, b)$
- (10) $\text{Par}(a, b; c, d) \rightarrow a \neq c \& a \neq d \& b \neq c \& b \neq d$
- (11) $\text{Par}(a, b; c, d) \leftrightarrow a \neq b \& c \neq d \& (Ex)(Eu)((R(a, x, u) \vee x = a) \& (R(b, x, u) \vee x = b) \& (R(x, u, c) \vee u = c) \& (R(x, u, d) \vee u = d))$

Für den Beweis der Implikation von rechts nach links hat man zu zeigen, dass auf einer Geraden a, b , mindestens fünf verschiedene Punkte liegen, was mit Hilfe der Axiome A1–A6 gelingt.

$$(12) \quad \text{Par}(a, b; c, d) \rightarrow (x)((R(a, x, c) \vee x = a) \& (R(b, x, c) \vee x = b) \rightarrow R(x, c, d))$$

$$(13) \quad \text{Par}(a, b; c, d) \& \text{Koll}(a, b, e) \& b \neq e \rightarrow \text{Par}(b, e; c, d)$$

und daraus insbesondere

$$(14) \quad \text{Par}(a, b; c, d) \rightarrow \neg \text{Koll}(a, b, c);$$

ferner

- (15) $\text{Par}(a, b; c, d) \& \text{Koll}(a, b, e) \rightarrow \neg \text{Koll}(c, d, e)$
- (16) $\neg \text{Koll}(a, b, c) \rightarrow (Ex)\text{Par}(a, b; c, x)$
- (17) $\text{Par}(a, b; c, d) \& \text{Par}(a, b; c, e) \rightarrow \text{Koll}(c, d, e)$
- (18) $\text{Par}(a, b; c, d) \& \text{Par}(a, b; e, f) \rightarrow \text{Par}(c, d; e, f) \vee (\text{Koll}(e, c, d) \& \text{Koll}(f, c, d)).$

An den Begriff des Parallelismus knüpft sich noch der der Vektor-

gleichheit: „ a, b und c, d sind die Gegenseiten eines Parallelogramms“:

DEFINITION 3. $\text{Pag}(a b; c, d) \leftrightarrow \text{Par}(a, b, c; d) \& \text{Par}(a, c; b, d)$

Man kann hiermit beweisen:

$$(19) \text{Pag}(a, b; c, d) \rightarrow \text{Pag}(c, d; a, b) \& \text{Pag}(a, c; b, d)$$

$$(20) \text{Pag}(a, b; c, d) \& \text{Pag}(a, b; c, e) \rightarrow d = e$$

$$(21) \text{Pag}(a, b; c, d) \rightarrow \neg \text{Koll}(a, b, c).$$

Für den Beweis des Existenzsatzes

$$(22) \neg \text{Koll}(a, b, c) \rightarrow (Ex)(\text{Pag}(a, b; c, x))$$

bedarf es noch eines weiteren Axioms:

$$A8 \ R(a, b, c) \rightarrow (Ex)(R(a, c, x) \& R(c, b, x)).$$

Mit Hilfe dieses Axioms ist generell beweisbar, dass zwei verschiedene, nicht parallele Geraden einen Schnittpunkt besitzen:

$$(23) \neg \text{Koll}(a, b, c) \& \neg \text{Par}(a, b; c, d) \rightarrow (Ex)(\text{Koll}(a, b, x) \& \text{Koll}(c, d, x)). -$$

Ob sich im Ganzen eine übersichtliche Axiomatik mit dem Grundbegriff R erreichen lässt, bleibe dahingestellt. Wir begnügen uns hier damit, Definitionen für die wesentlichen weiteren Begriffe aufzustellen. Für diese lässt sich immerhin eine gewisse Übersichtlichkeit erreichen.

An die Figur des Parallelogramms knüpfen sich die folgenden zwei verschiedenen Definitionen der Beziehung „ a ist Mittelpunkt der Strecke b, c “:

$$\text{DEFINITION 4}_1 \ Mp_1(a; b, c) \leftrightarrow (Ex)(Ey)(\text{Pag}(b, x; y, c) \& \text{Koll}(a, b, c) \& \text{Koll}(a, x, y))$$

$$\text{DEFINITION 4}_2 \ Mp_2(a; b, c) \leftrightarrow (Ex)(Ey)(\text{Pag}(x, y; a, b) \& \text{Pag}(x, y; c, a)).$$

Im Sinne der zweiten Definition kann man die Möglichkeit der Verdoppelung einer Strecke beweisen:

$$(24) a \neq b \rightarrow (Eu)Mp_2(a; b, u).$$

Die Existenz des Mittelpunktes einer Strecke im Sinne der Df. 4₁, d. h.

$$(25) b \neq c \rightarrow (Eu)Mp_1(u; b, c),$$

lässt sich beweisen, wenn man noch das Axiom hinzunimmt:

A9 $\text{Par}(a, b; c, d) \& \text{Par}(a, c; b, d) \rightarrow \neg \text{Par}(a, d; b, c).$

(Im Parallelogramm schneiden sich die Diagonalen)

Durch Spezialisierung der zur Definition von Mp_1 gehörigen Figur erhalten wir eine Definition der Beziehung „ a, b, c bilden ein gleichschenkliges Dreieck mit der Spitze in a “:

DEFINITION 51. $Ist_1(a; b, c) \leftrightarrow (Eu)(Ev)(\text{Pag}(a, b; c, v) \& R(a, u, b) \& R(a, u, c) \& R(b, u, v)).$

Mit Hilfe von Mp_1 und Ist_1 können wir den Pieri'schen Grundbegriff: „ a hat von b und c gleichen Abstand“ definieren:

DEFINITION 6. $Is_1(a; b, c) \leftrightarrow b = c \vee Mp_1(a; b, c) \vee Ist_1(a; b, c).$

Eine andere Art der Definition des Begriffes Is beruht auf der Verwendung der Symmetrie. Hierzu dient folgender Hilfsbegriff: „ a, b, c, d, e bilden ein „normales“ Quintupel“:

DEFINITION 7. $Qn(a, b, c, d, e) \leftrightarrow R(a, c, b) \& R(a, d, b) \& R(a, e, c) \& R(a, e, d) \& R(b, e, c) \& c \neq d.$

Mit Hilfe von Qn erhalten wir eine weitere Art der Definition für Mp und Ist :

DEFINITION $\begin{cases} 43. & Mp_3(a; b, c) \leftrightarrow (Ex)(Ey)Qn(x, y, b, c, a) \\ 52. & Ist_2(a; b, c) \leftrightarrow (Ex)(Ey)Qn(a, x, b, c, y), \end{cases}$

aus denen sich Is_2 entsprechend wie Is_1 definieren lässt.

Ferner schliesst sich hieran noch die Definition der Spiegelbildlichkeit von Punkten a, b in Bezug auf eine Gerade c, d :

DEFINITION 8. $\text{Sym}(a, b; c, d) \leftrightarrow c \neq d \& (Ex)(Ey)(Ez)(\text{Koll}(x, c, d) \& \text{Koll}(y, c, d) \& Qn(x, y, a, b, z)).$ —

Für die Definition der Streckenkongruenz brauchen wir schliesslich noch den Begriff der gleichsinnigen Kongruenz auf einer Geraden: „die Strecken a, b und c, d sind kollinear, kongruent und gleichgerichtet“:

DEFINITION 91. $Lg_1(a, b; c, d) \leftrightarrow \text{Koll}(a, b, c) \& (Ex)(Ey)(\text{Pag}(a, x; b, y) \& \text{Pag}(c, x; d, y)),$

oder auch:

DEFINITION 92. $Lg_2(a, b; c, d) \leftrightarrow \text{Koll}(a, b, c) \& a \neq b \& (Ex)(Mp(x; b, c) \& Mp(x; a, d)) \vee (a = d \& Mp(a; b, c)) \vee (b = c \& Mp(b; a, d)),$

(wobei für $M\phi$ eine der drei obigen Definitionen genommen werden kann. Nunmehr kann im Ganzen (mit jeder der beiden Definitionen von Lg) die Streckenkongruenz definiert werden:

DEFINITION 10. $Kg(a, b; c, d) \leftrightarrow Lg(a, b; c, d) \vee Lg(a, b; d, c) \vee$
 $\vee (a = b \ \& \ Is_1(a; b, d)) \vee (Ex)(Pag(a, b; c, x) \ \& \ Is_1(c; x, d)).$

Durch eine Definition analog derjenigen von Lg_2 kann man auch die Kongruenz von Winkeln mit gleichem Scheitelpunkt als sechsstellige Beziehung einführen, nachdem man vorher den Begriff der Winkelhalbierenden eingeführt hat: „ $d(\neq a)$ liegt auf der Halbierenden des Winkels $b \ a \ c'$ “:

DEFINITION 11. $Wh(a, d; b, c) \leftrightarrow \neg Koll(a, b, c) \ \&$
 $\ \& \ (Ex)(Ey)(Ez)(Koll(a, c, x) \ \& \ Koll(a, d, y) \ \& \ Qn(a, y, b, x, z)).$

In Anbetracht des sehr zusammengesetzten Charakters dieser Kongruenzbeziehung Kg wird man in der Axiomatisierung die Gesetze über Kg auf solche der als Bestandteile des definierenden Ausdrucks auftretenden Begriffe zurückführen. Dabei bestehen auf Grund der Mehrheit der Definitionen von $M\phi$, Ist , Is Alternativen in Hinsicht darauf, ob man in stärkerem Masse die Beziehungen des Parallelismus oder die der Symmetrie heranzieht. Auf jeden Fall dürfte das Axiom der Vektorgeometrie

A10 $Pag(a, b; p, q) \ \& \ Pag(b, c; q, r) \rightarrow Pag(a, c; p, r) \vee$
 $\vee (Koll(a, c, p) \ \& \ Koll(a, c, r))$

oder ein gleichwertiges zweckmässig sein. Im Ganzen könnte man sich hierbei als Ziel setzen, das in der euklidischen Planimetrie vorliegende Zusammenspiel von Parallelismus und Spiegelung auf eine möglichst symmetrische Art zur Darstellung zu bringen.

Was endlich die Zwischenbeziehung betrifft, so ist die Figur für die Definition der Beziehung „ a liegt zwischen b und c “ schon als Bestandteil in derjenigen von Qn enthalten. Nämlich wir können definieren:

DEFINITION 12. $Zw(a; b, c) \leftrightarrow (Ex)(R(b, a, x) \ \& \ R(c, a, x) \ \& \ R(b, x, c)).$

Für diesen Begriff sind zunächst beweisbar:

$$(26) \quad \neg Zw(a; b, b)$$

$$(27) \quad Zw(a; b, c) \rightarrow Zw(a; c, b)$$

$$(28) \quad Zw(a; b, c) \rightarrow Koll(a, b, c)$$

und ferner mit Benutzung von A5, A6 und A8

$$(29) \quad a \neq b \rightarrow (Ex)Zw(x; a, b) \& (Ex)Zw(b; a, x).$$

Für die Gewinnung der weiteren Eigenschaften des Zwischenbegriffes können die folgenden Axiome dienen:

$$A11 \quad R(a, b, c) \& R(a, b, d) \& R(c, a, d) \& R(e, c, b) \rightarrow \neg R(b, e, d)$$

$$A12 \quad R(a, b, d) \& R(d, b, c) \& a \neq c \rightarrow Zw(a; b, c) \vee Zw(b; a, c) \vee Zw(c; a, b)$$

$$A13 \quad Zw(a; b, c) \& Zw(b; a, d) \rightarrow Zw(a; c, d)$$

$$A14 \quad R(a, b, d) \& R(d, b, c) \& R(a, c, e) \& Zw(d; a, e) \rightarrow Zw(b; a, c)$$

Aus diesem Axiom kann man in einigen Schritten den allgemeineren Satz gewinnen:

$$(30) \quad Zw(b; a, c) \& Koll(a, d, e) \& Par(b, d; c, e) \rightarrow Zw(d; a, e).$$

Dieses gelingt mit Verwendung des Satzes

$$(31) \quad R(a, b, e) \& R(e, b, c) \& R(b, a, d) \& R(b, c, f) \& R(b, e, d) \& R(b, e, f) \& Zw(b; a, c) \rightarrow Zw(e; d, f),$$

welcher sich aus dem vorhin erwähnten Axiom A10 ableiten lässt.

Mit Hilfe von (30) und dem Axiom A13 lässt sich beweisen:

$$(32) \quad \neg Koll(a, b, c) \& Zw(b; a, d) \& Zw(e; b, c) \rightarrow (Ex)(Koll(e, d, x) \& Zw(x; a, c)),$$

dh. das Axiom von Pasch in der engeren Veblen'schen Fassung.—

Anschliessend sei noch die folgende Definition von Kg mittels der Begriffe Is und Zw erwähnt, welche auf einer Konstruktion von Euklid beruht:

$$\text{DEFINITION 13.} \quad Kg^*(a, b; c, d) \leftrightarrow (Ex)(Ey)(Ez)(Is(x, a; c) \& Zw(y; a, x) \& Zw(z; c, x) \& Is(a; b, y) \& Is(c; d, z) \& Is(x; y, z)).$$

(Für Is kann hier nach Belieben Is_1 oder Is_2 genommen werden.)

Von einer Axiomatik wie der hier geschilderten, bei der die Kollinearität und die Zwischenbeziehung mit der Orthogonalität verkoppelt wird, kann man freilich nicht verlangen, dass sie eine Absonderung der Axiome

des Linearen liefert. Ferner ist die Anlage hier von vornherein im Hinblick auf die Planimetrie beschränkt, da die Definition der Kollinearität im Mehrdimensionalen nicht mehr anwendbar ist. Auch die Beschränkung auf die euklidische Geometrie wird schon an früher Stelle eingeführt. Andererseits kann diese Axiomatisierung sich besonders dafür eignen, die grosse Einfachheit und Eleganz der Gesetzlichkeit der euklidischen Planimetrie hervortreten zu lassen.

WHAT IS ELEMENTARY GEOMETRY?

ALFRED TARSKI

*Institute for Basic Research in Science,
University of California, Berkeley, California, U.S.A.*

In colloquial language the term *elementary geometry* is used loosely to refer to the body of notions and theorems which, following the tradition of Euclid's *Elements*, form the subject matter of geometry courses in secondary schools. Thus the term has no well determined meaning and can be subjected to various interpretations. If we wish to make elementary geometry a topic of metamathematical investigation and to obtain exact results (not within, but) about this discipline, then a choice of a definite interpretation becomes necessary. In fact, we have then to describe precisely which sentences can be formulated in elementary geometry and which among them can be recognized as valid; in other words, we have to determine the means of expression and proof with which the discipline is provided.

In this paper we shall primarily concern ourselves with a conception of elementary geometry which can roughly be described as follows: *we regard as elementary that part of Euclidean geometry which can be formulated and established without the help of any set-theoretical devices.*¹

More precisely, elementary geometry is conceived here as a theory with standard formalization in the sense of [9].² It is formalized within ele-

¹ The paper was prepared for publication while the author was working on a research project in the foundations of mathematics sponsored by the U.S. National Science Foundation.

² One of the main purposes of this paper is to exhibit the significance of notions and methods of modern logic and metamathematics for the study of the foundations of geometry. For logical and metamathematical notions involved in the discussion consult [8] and [9] (see the bibliography at the end of the paper). The main metamathematical result upon which the discussion is based was established in [7]. For algebraic notions and results consult [11].

Several articles in this volume are related to the present paper in methods and results. This applies in the first place to Scott [5] and Szmielew [6], and to some extent also to Robinson [3].

mentary logic, i.e., first-order predicate calculus. All the variables x, y, z, \dots occurring in this theory are assumed to range over elements of a fixed set; the elements are referred to as points, and the set as the space. The logical constants of the theory are (i) the sentential connectives — the negation symbol \neg , the implication symbol \rightarrow , the disjunction symbol \vee , and the conjunction symbol \wedge ; (ii) the quantifiers — the universal quantifier Λ and the existential quantifier \mathbf{V} ; and (iii) two special binary predicates — the identity symbol $=$ and the diversity symbol \neq . As non-logical constants (primitive symbols of the theory) we could choose any predicates denoting certain relations among points in terms of which all geometrical notions are known to be definable. Actually we pick two predicates for this purpose: the ternary predicate β used to denote the betweenness relation and the quaternary predicate δ used to denote the equidistance relation; the formula $\beta(xyz)$ is read *y lies between x and z* (the case when y coincides with x or z not being excluded), while $\delta(xyzu)$ is read *x is as distant from y as z is from u*.

Thus, in our formalization of elementary geometry, only points are treated as individuals and are represented by (first-order) variables. Since elementary geometry has no set-theoretical basis, its formalization does not provide for variables of higher orders and no symbols are available to represent or denote geometrical figures (point sets), classes of geometrical figures, etc. It should be clear that, nevertheless, we are able to express in our symbolism all the results which can be found in textbooks of elementary geometry and which are formulated there in terms referring to various special classes of geometrical figures, such as the straight lines, the circles, the segments, the triangles, the quadrangles, and, more generally, the polygons with a fixed number of vertices, as well as to certain relations between geometrical figures in these classes, such as congruence and similarity. This is primarily a consequence of the fact that, in each of the classes just mentioned, every geometrical figure is determined by a fixed finite number of points. For instance, instead of saying that a point z lies on the straight line through the points x and y , we can state that either $\beta(xyz)$ or $\beta(yzx)$ or $\beta(zxy)$ holds; instead of saying that two segments with the end-points x, y and x', y' are congruent, we simply state that $\delta(xy x' y')$.³

³ In various formalizations of geometry (whether elementary or not) which are known from the literature, and in particular in all those which follow the lines of [1], not only points but also certain special geometrical figures are treated as

A sentence formulated in our symbolism is regarded as valid if it follows (semantically) from sentences adopted as axioms, i.e., if it holds in every mathematical structure in which all the axioms hold. In the present case, by virtue of the completeness theorem for elementary logic, this amounts to saying that a sentence is valid if it is derivable from the axioms by means of some familiar rules of inference. To obtain an appropriate set of axioms, we start with an axiom system which is known to provide an adequate basis for the whole of Euclidean geometry and contains β and δ as the only non-logical constants. Usually the only non-elementary sentence in such a system is the continuity axiom, which contains second-order variables X, Y, \dots ranging over arbitrary point sets (in addition to first-order variables x, y, \dots ranging over points) and also an additional logical constant, the membership symbol \in denoting the membership relation between points and point sets. The continuity axiom can be formulated, e.g., as follows:

$$\begin{aligned} \wedge XY \{ \forall z \wedge xy [x \in X \wedge y \in Y \rightarrow \beta(zxy)] \\ \rightarrow \forall u \wedge xy [x \in X \wedge y \in Y \rightarrow \beta(xuy)] \}. \end{aligned}$$

We remove this axiom from the system and replace it by the infinite collection of all elementary continuity axioms, i.e., roughly, by all the sentences which are obtained from the non-elementary axiom if $x \in X$ is replaced by an arbitrary elementary formula in which x occurs free, and $y \in Y$ by an arbitrary elementary formula in which y occurs free. To fix the ideas, we restrict ourselves in what follows to the two-dimensional

individuals and are represented by first-order variables; usually the only figures treated this way are straight lines, planes, and, more generally, linear subspaces. The set-theoretical relations of membership and inclusion, between a point and a special geometrical figure or between two such figures, are replaced by the geometrical relation of incidence, and the symbol denoting this relation is included in the list of primitive symbols of geometry. All other geometrical figures are treated as point sets and can be represented by second-order variables (assuming that the system of geometry discussed is provided with a set-theoretical basis). This approach has some advantages for restricted purposes of projective geometry; in fact, it facilitates the development of projective geometry by yielding a convenient formulation of the duality principle, and leads to a subsumption of this geometry under the algebraic theory of lattices. In other branches of geometry an analogous procedure can hardly be justified; the non-uniform treatment of geometrical figures seems to be intrinsically unnatural, obscures the logical structure of the foundations of geometry, and leads to some complications in the development of this discipline (by necessitating, e.g., a distinction between a straight line and the set of all points on this line).

elementary geometry and quote explicitly a simple axiom system obtained in the way just described. The system consists of twelve individual axioms, A1–A2, and the infinite collection of all elementary continuity axioms, A13.

A1 [IDENTITY AXIOM FOR BETWEENNESS].

$$\bigwedge xy[\beta(xyx) \rightarrow (x = y)]$$

A2 [TRANSITIVITY AXIOM FOR BETWEENNESS].

$$\bigwedge xyzu[\beta(xyu) \wedge \beta(yzu) \rightarrow \beta(xyz)]$$

A3 [CONNECTIVITY AXIOM FOR BETWEENNESS].

$$\bigwedge xyzu[\beta(xyz) \wedge \beta(xyu) \wedge (x \neq y) \rightarrow \beta(xzu) \vee \beta(xuz)]$$

A4 [REFLEXIVITY AXIOM FOR EQUIDISTANCE].

$$\bigwedge xy[\delta(xyyx)]$$

A5 [IDENTITY AXIOM FOR EQUIDISTANCE].

$$\bigwedge xyz[\delta(xyzz) \rightarrow (x = y)]$$

A6 [TRANSITIVITY AXIOM FOR EQUIDISTANCE].

$$\bigwedge xyzuvw[\delta(xyzu) \wedge \delta(xyvw) \rightarrow \delta(zuvw)]$$

A7 [PASCH'S AXIOM].

$$\bigwedge txyzu \vee v[\beta(xtu) \wedge \beta(yuz) \rightarrow \beta(xvy) \wedge \beta(ztv)]$$

A8 [EUCLID'S AXIOM].

$$\bigwedge txyzu \vee vw[\beta(xut) \wedge \beta(yuz) \wedge (x \neq u) \rightarrow \beta(xzv) \wedge \beta(xyw) \wedge \beta(vtw)]$$

A9 (FIVE-SEGMENT AXIOM).

$$\begin{aligned} \bigwedge xx'yy'zz'u'u'[\delta(xyx'y') \wedge \delta(yzy'z') \wedge \delta(xux'u') \wedge \delta(yuy'u') \\ \wedge \beta(xyz) \wedge \beta(x'y'z') \wedge (x \neq y) \rightarrow \delta(zuz'u')] \end{aligned}$$

A10 (AXIOM OF SEGMENT CONSTRUCTION).

$$\bigwedge xyuv \vee z[\beta(xyz) \wedge \delta(yzuv)]$$

A11 (LOWER DIMENSION AXIOM).

$$\bigvee xyz[\neg\beta(xyz) \wedge \neg\beta(yzx) \wedge \neg\beta(zxy)]$$

A12 (UPPER DIMENSION AXIOM).

$$\begin{aligned} \bigwedge xyzuv[\delta(xuxv) \wedge \delta(yuyv) \wedge \delta(zuzv) \wedge (u \neq v) \\ \rightarrow \beta(xyz) \vee \beta(yzx) \vee \beta(zxy)] \end{aligned}$$

A13 [ELEMENTARY CONTINUITY AXIOMS]. *All sentences of the form*

$$\Lambda vw \dots \{ \forall z \Lambda xy [\varphi \wedge \psi \rightarrow \beta(zxy)] \rightarrow \forall u \Lambda xy [\varphi \wedge \psi \rightarrow \beta(xuy)] \}$$

where φ stands for any formula in which the variables x, v, w, \dots , but neither y nor z nor u , occur free, and similarly for ψ , with x and y interchanged.

Elementary geometry based upon the axioms just listed will be denoted by \mathcal{E}_2 . In Theorems 1–4 below we state fundamental metamathematical properties of this theory.⁴

First we deal with the *representation problem* for \mathcal{E}_2 , i.e., with the problem of characterizing all models of this theory. By a model of \mathcal{E}_2 we understand a system $\mathfrak{M} = \langle A, B, D \rangle$ such that (i) A is an arbitrary non-empty set, and B and D are respectively a ternary and a quaternary relation among elements of A ; (ii) all the axioms of \mathcal{E}_2 prove to hold in \mathfrak{M} if all the variables are assumed to range over elements of A , and the constants β and δ are understood to denote the relations B and D , respectively.

The most familiar examples of models of \mathcal{E}_2 (and ones which can easily be handled by algorithmic methods) are certain Cartesian spaces over ordered fields. We assume known under what conditions a system $\mathfrak{F} = \langle F, +, \cdot, \leq \rangle$ (where F is a set, $+$ and \cdot are binary operations under which F is closed, and \leq is a binary relation between elements of F) is referred to as an ordered field and how the symbols $0, x - y, x^2$ are defined for ordered fields. An ordered field \mathfrak{F} will be called Euclidean if every non-negative element in F is a square; it is called real closed if it is Euclidean and if every polynomial of an odd degree with coefficients in F has a zero in F . Consider the set $A_{\mathfrak{F}} = F \times F$ of all ordered couples

⁴ A brief discussion of the theory \mathcal{E}_2 and its metamathematical properties was given in [7], pp. 43 ff. A detailed development (based upon the results of [7]) can be found in [4] — where, however, the underlying system of elementary geometry differs from the one discussed in this paper in its logical structure, primitive symbols, and axioms.

The axiom system for \mathcal{E}_2 quoted in the text above is a simplified version of the system in [7], pp. 55 f. The simplification consists primarily in the omission of several superfluous axioms. The proof that those superfluous axioms are actually derivable from the remaining ones was obtained by Eva Kallin, Scott Taylor, and the author in connection with a course in the foundations of geometry given by the author at the University of California, Berkeley, during the academic year 1956–57.

$x = \langle x_1, x_2 \rangle$ with x_1 and x_2 in F . We define the relations $B_{\mathfrak{F}}$ and $D_{\mathfrak{F}}$ among such couples by means of the following stipulations:

$B_{\mathfrak{F}}(xyz)$ if and only if $(x_1 - y_1) \cdot (y_2 - z_2) = (x_2 - y_2) \cdot (y_1 - z_1)$,

$0 \leq (x_1 - y_1) \cdot (y_1 - z_1)$, and $0 \leq (x_2 - y_2) \cdot (y_2 - z_2)$;

$D_{\mathfrak{F}}(xyzu)$ if and only if $(x_1 - y_1)^2 + (x_2 - y_2)^2 = (z_1 - u_1)^2 + (z_2 - u_2)^2$.

The system $\mathcal{C}_2(\mathfrak{F}) = \langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ is called the (two-dimensional) Cartesian space over \mathfrak{F} . If in particular we take for \mathfrak{F} the ordered field \mathbb{R} of real numbers, we obtain the ordinary (two-dimensional) analytic space $\mathcal{C}_2(\mathbb{R})$.⁵

THEOREM 1 (REPRESENTATION THEOREM). *For \mathfrak{M} to be a model of \mathcal{E}_2 it is necessary and sufficient that \mathfrak{M} be isomorphic with the Cartesian space $\mathcal{C}_2(\mathfrak{F})$ over some real closed field \mathfrak{F} .*

PROOF (in outline). It is well known that all the axioms of \mathcal{E}_2 hold in $\mathcal{C}_2(\mathbb{R})$ and that therefore $\mathcal{C}_2(\mathbb{R})$ is a model of \mathcal{E}_2 . By a fundamental result in [7], every real closed field \mathfrak{F} is elementarily equivalent with the field \mathbb{R} , i.e., every elementary (first-order) sentence which holds in one of these two fields holds also in the other. Consequently every Cartesian space $\mathcal{C}_2(\mathfrak{F})$ over a real closed field \mathfrak{F} is elementarily equivalent with $\mathcal{C}_2(\mathbb{R})$ and hence is a model of \mathcal{E}_2 ; this clearly applies to all systems \mathfrak{M} isomorphic with $\mathcal{C}_2(\mathfrak{F})$ as well.

To prove the theorem in the opposite direction, we apply methods and results of the elementary geometrical theory of proportions, which has been developed in the literature on several occasions (see, e.g., [1], pp. 51 ff.). Consider a model $\mathfrak{M} = \langle A, B, D \rangle$ of \mathcal{E}_2 ; let z and u be any two distinct points of A , and F be the straight line through z and u , i.e., the set of all points x such that $B(zux)$ or $B(uxz)$ or $B(xzu)$. Applying some familiar geometrical constructions, we define the operations $+$ and \cdot on, and the relation \leq between, any two points x and y in F . Thus we say that $x \leq y$ if either $x = y$ or else $B(xzu)$ and not $B(yxu)$ or, finally,

⁵ All the results in this paper extend (with obvious changes) to the n -dimensional case for any positive integer n . To obtain an axiom system for \mathcal{E}_n we have to modify the two dimension axioms, A11 and A12, leaving the remaining axioms unchanged; by a result in [5], A11 and A12 can be replaced by any sentence formulated in the symbolism of \mathcal{E}_n which holds in the ordinary n -dimensional analytic space but not in any m -dimensional analytic space for $m \neq n$. In constructing algebraic models for n -dimensional geometries we use ordered abelian groups instead of ordered fields.

$B(zxy)$ and not $B(xzu)$; $x + y$ is defined as the unique point v in F such that $D(zxyv)$ and either $z \leq x$ and $y \leq v$ or else $x \leq z$ and $v \leq y$. The definition of $x \cdot y$ is more involved; it refers to some points outside of F and is essentially based upon the properties of parallel lines. Using exclusively axioms A1–A12 we show that $\mathfrak{F} = \langle F, +, \cdot, \leq \rangle$ is an ordered field; with the help of A13 we arrive at the conclusion that \mathfrak{F} is actually a real closed field. By considering a straight line G perpendicular to F at the point z , we introduce a rectangular coordinate system in \mathfrak{M} and we establish a one-to-one correspondence between points x, y, \dots in A and ordered couples of their coordinates $\bar{x} = \langle x_1, x_2 \rangle$, $\bar{y} = \langle y_1, y_2 \rangle$, \dots in $F \times F$. With the help of the Pythagorean theorem (which proves to be valid in \mathcal{E}_2) we show that the formula

$$D(xyst)$$

holds for any given points x, y, \dots in A if and only if the formula

$$D_{\mathfrak{F}}(\bar{x}\bar{y}\bar{s}\bar{t})$$

holds for the correlated couples of coordinates $\bar{x} = \langle x_1, x_2 \rangle$, $\bar{y} = \langle y_1, y_2 \rangle$, \dots in $F \times F$, i.e., if

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 = (s_1 - t_1)^2 + (s_2 - t_2)^2;$$

an analogous conclusion is obtained for $B(xys)$. Consequently, the systems \mathfrak{M} and $\mathfrak{C}_2(\mathfrak{F})$ are isomorphic, which completes the proof.

We turn to the *completeness problem* for \mathcal{E}_2 . A theory is called complete if every sentence σ (formulated in the symbolism of the theory) holds either in every model of this theory or in no such model. For theories with standard formalization this definition can be put in several other equivalent forms; we can say, e.g., that a theory is complete if, for every sentence σ , either σ or $\neg\sigma$ is valid, or if any two models of the theory are elementarily equivalent. A theory is called consistent if it has at least one model; here, again, several equivalent formulations are known. If there is a model \mathfrak{M} such that a sentence holds in \mathfrak{M} if and only if it is valid in the given theory, then the theory is clearly both complete and consistent, and conversely. The solution of the completeness problem for \mathcal{E}_2 is given in the following

- THEOREM 2 (COMPLETENESS THEOREM). (i) *A sentence formulated in \mathcal{E}_2 is valid if and only if it holds in $\mathfrak{C}_2(\mathfrak{R})$;*
(ii) *the theory \mathcal{E}_2 is complete (and consistent).*

Part (i) of this theorem follows from Theorem 1 and from a fundamental result in [7] which was applied in the proof of Theorem 1; (ii) is an immediate consequence of (i).

The next problem which will be discussed here is the *decision problem* for \mathcal{E}_2 . It is the problem of the existence of a mechanical method which enables us in each particular case to decide whether or not a given sentence formulated in \mathcal{E}_2 is valid. The solution of this problem is again positive:

THEOREM 3 (DECISION THEOREM). *The theory \mathcal{E}_2 is decidable.*

In fact, \mathcal{E}_2 is complete by Theorem 2 and is axiomatizable by its very description (i.e., it has an axiom system such that we can always decide whether a given sentence is an axiom). It is known, however, that every complete and axiomatizable theory with standard formalization is decidable (cf., e.g., [9], p. 14), and therefore \mathcal{E}_2 is decidable. By analyzing the discussion in [7] we can actually obtain a decision method for \mathcal{E}_2 .

The last metamathematical problem to be discussed for \mathcal{E}_2 is the *problem of finite axiomatizability*. From the description of \mathcal{E}_2 we see that this theory has an axiom system consisting of finitely many individual axioms and of an infinite collection of axioms falling under a single axiom schema. This axiom schema (which is the symbolic expression occurring in A13) can be slightly modified so as to form a single sentence in the system of predicate calculus with free variable first-order predicates, and all the particular axioms of the infinite collection can be obtained from this sentence by substitution. We briefly describe the whole situation by saying that the theory \mathcal{E}_2 is "almost finitely axiomatizable", and we now ask the question whether \mathcal{E}_2 is finitely axiomatizable in the strict sense, i.e., whether the original axiom system can be replaced by an equivalent finite system of sentences formulated in \mathcal{E}_2 . The answer is negative:

THEOREM 4 (NON-FINITIZABILITY THEOREM). *The theory \mathcal{E}_2 is not finitely axiomatizable.*

PROOF (in outline). From the proof of Theorem 1 it is seen that the infinite collection of axioms A13 can be equivalently replaced by an infinite sequence of sentences S_0, \dots, S_n, \dots ; S_0 states that the ordered field \mathfrak{F} constructed in the proof of Theorem 1 is Euclidean, and S_n for $n > 0$ expresses the fact that in this field every polynomial of degree $2n + 1$ has a zero. For every prime number p we can easily construct an ordered

field \mathfrak{F}_p in which every polynomial of an odd degree $2n + 1 < p$ has a zero while some polynomial of degree p has no zero; consequently, if $2m + 1 = p$ is a prime, then all the axioms A1–A12 and S_n with $n < m$ hold in $\mathfrak{C}_2(\mathfrak{F}_p)$ while S_m does not hold. This implies immediately that the infinite axiom system A1, ..., A12, S_0 , ..., S_n , ... has no finite subsystem from which all the axioms of the system follow. Hence by a simple argument we conclude that, more generally, there is no finite axiom system which is equivalent with the original axiom system for \mathcal{E}_2 .

From the proof just outlined we see that \mathcal{E}_2 can be based upon an axiom system A1, ..., A12, S_0 , ..., S_n , ... in which (as opposed to the original axiom system) each axiom can be put in the form of either a universal sentence or an existential sentence or a universal-existential sentence; i.e., each axiom is either of the form

$$\bigwedge xy \dots (\varphi)$$

or else of the form

$$\bigvee uv \dots (\varphi)$$

or, finally, of the form

$$\bigwedge xy \dots \bigvee uv \dots (\varphi)$$

where φ is a formula without quantifiers. A rather obvious consequence of this structural property of the axioms is the fact that the union of a chain (or of a directed family) of models of \mathcal{E}_2 is again a model of \mathcal{E}_2 . This consequence can also be derived directly from the proof of Theorem 1.

The conception of elementary geometry with which we have been concerned so far is certainly not the only feasible one. In what follows we shall discuss briefly two other possible interpretations of the term "elementary geometry"; they will be embodied in two different formalized theories, \mathcal{E}_2' and \mathcal{E}_2'' .

The theory \mathcal{E}_2' is obtained by supplementing the logical base of \mathcal{E}_2 with a small fragment of set theory. Specifically, we include in the symbolism of \mathcal{E}_2' new variables X, Y, \dots assumed to range over arbitrary finite sets of points (or, what in this case amounts essentially to the same, over arbitrary finite sequences of points); we also include a new logical constant, the membership symbol \in , to denote the membership relation between points and finite point sets. As axioms for \mathcal{E}_2' we again choose A1–A13; it should be noticed, however, that the collection of axiom A13

is now more comprehensive than in the case of \mathcal{E}_2 since φ and ψ stand for arbitrary formulas constructed in the symbolism of \mathcal{E}_2' . In consequence the theory \mathcal{E}_2' considerably exceeds \mathcal{E}_2 in means of expression and power. In \mathcal{E}_2' we can formulate and study various notions which are traditionally discussed in textbooks of elementary geometry but which cannot be expressed in \mathcal{E}_2 ; e.g., the notions of a polygon with arbitrarily many vertices, and of the circumference and the area of a circle.

As regards metamathematical problems which have been discussed and solved for \mathcal{E}_2 in Theorems 1-4, three of them — the problems of representation, completeness, and finite axiomatizability — are still open when referred to \mathcal{E}_2' . In particular, we do not know any simple characterization of all models of \mathcal{E}_2' , nor, do we know whether any two such models are equivalent with respect to all sentences formulated in \mathcal{E}_2' . (When speaking of models of \mathcal{E}_2' we mean exclusively the so-called standard models; i.e., when deciding whether a sentence σ formulated in \mathcal{E}_2' holds in a given model, we assume that the variables x, y, \dots occurring in σ range over all elements of a set, the variables X, Y, \dots range over all finite subsets of this set, and \in is always understood to denote the membership relation). The Archimedean postulate can be formulated and proves to be valid in \mathcal{E}_2' . Hence, by Theorem 1, every model of \mathcal{E}_2' is isomorphic with a Cartesian space $\mathcal{C}_2(\mathfrak{F})$ over some Archimedean real closed field \mathfrak{F} . There are, however, Archimedean real closed fields \mathfrak{F} such that $\mathcal{C}_2(\mathfrak{F})$ is not a model of \mathcal{E}_2' ; e.g., the field of real algebraic numbers is of this kind. A consequence of the Archimedean postulate is that every model of \mathcal{E}_2' has at most the power of the continuum (while, if only by virtue of Theorem 1, \mathcal{E}_2 has models with arbitrary infinite powers). In fact, \mathcal{E}_2' has models which have exactly the power of the continuum, e.g., $\mathcal{C}_2(\mathbb{R})$, but it can also be shown to have denumerable models. Thus, although the theory \mathcal{E}_2' may prove to be complete, it certainly has non-isomorphic models and therefore is not categorical.⁶

Only the decision problem for \mathcal{E}_2' has found so far a definite solution:

⁶ These last remarks result from a general metamathematical theorem (an extension of the Skolem-Löwenheim theorem) which applies to all theories with the same logical structure as \mathcal{E}_2' , i.e., to all theories obtained from theories with standard formalization by including new variables ranging over arbitrary finite sets and a new logical constant, the membership symbol \in , and possibly by extending original axiom systems. By this general theorem, if \mathcal{T} is a theory of the class just described with at most β different symbols, and if a mathematical system \mathfrak{M} is a

THEOREM 5. *The theory \mathcal{E}_2' is undecidable, and so are all its consistent extensions.*

This follows from the fact that Peano's arithmetic is (relatively) interpretable in \mathcal{E}_2' ; cf. [9], pp. 31ff.

To obtain the theory \mathcal{E}_2'' we leave the symbolism of \mathcal{E}_2 unchanged but we weaken the axiom system of \mathcal{E}_2 . In fact, we replace the infinite collection of elementary continuity axioms, A13, by a single sentence, A13', which is a consequence of one of these axioms. The sentence expresses the fact that a segment which joins two points, one inside and one outside a given circle, always intersects the circle; symbolically:

$$\text{A13'}. \quad \bigwedge xyzx'z'u \vee y'[\delta(uxux') \wedge \delta(uzuz') \wedge \beta(uxz) \wedge \beta(xyz) \\ \rightarrow \delta(uyuy') \wedge \beta(x'y'z')]$$

As a consequence of the weakening of the axiom system, various sentences which are formulated and valid in \mathcal{E}_2 are no longer valid in \mathcal{E}_2'' . This applies in particular to existential theorems which cannot be established by means of so-called elementary geometrical constructions (using exclusively ruler and compass), e.g., to the theorem on the trisection of an arbitrary angle.

With regard to metamathematical problems discussed in this paper the situation in the case of \mathcal{E}_2'' is just opposite to that encountered in the case of \mathcal{E}_2' . The three problems which are open for \mathcal{E}_2' admit of simple solutions when referred to \mathcal{E}_2'' . In particular, the solution of the representation problem is given in the following

standard model of \mathcal{T} with an infinite power α , then \mathfrak{M} has subsystems with any infinite power γ , $\beta \leq \gamma \leq \alpha$, which are also standard models of \mathcal{T} . The proof of this theorem (recently found by the author) has not yet been published; it differs but slightly from the proof of the analogous theorem for the theories with standard formalization outlined in [10], pp. 92 f. In opposition to theories with standard formalization, some of the theories \mathcal{T} discussed in this footnote have models with an infinite power α and with any smaller, but with no larger, infinite power; an example is provided by the theory \mathcal{E}_2' for which α is the power of the continuum. In particular, some of the theories \mathcal{T} have exclusively denumerable models and in fact are categorical; this applies, e.g., to the theory obtained from Peano's arithmetic in exactly the same way in which \mathcal{E}_2' has been obtained from \mathcal{E}_2 . There are also theories \mathcal{T} which have models with arbitrary infinite powers; such is, e.g., the theory \mathcal{E}_2''' mentioned at the end of this paper.

THEOREM 6. *For \mathfrak{M} to be a model of \mathcal{E}_2'' it is necessary and sufficient that \mathfrak{M} be isomorphic with the Cartesian space $\mathcal{C}_2(\mathfrak{F})$ over some Euclidean field \mathfrak{F} .*

This theorem is essentially known from the literature. The sufficiency of the condition can be checked directly; the necessity can be established with the help of the elementary geometrical theory of proportions (cf. the proof of Theorem 1).

Using Theorem 6 we easily show that the theory \mathcal{E}_2'' is incomplete, and from the description of \mathcal{E}_2'' we see at once that this theory is finitely axiomatizable.

On the other hand, the decision problem for \mathcal{E}_2'' remains open and presumably is difficult. In the light of the results in [2] it seems likely that the solution of this problem is negative; the author would risk the (much stronger) conjecture that no finitely axiomatizable subtheory of \mathcal{E}_2 is decidable. If we agree to refer to an elementary geometrical sentence (i.e., a sentence formulated in \mathcal{E}_2) as valid if it is valid in \mathcal{E}_2 , and as elementarily provable if it is valid in \mathcal{E}_2'' , then the situation can be described as follows: *we know a general mechanical method for deciding whether a given elementary geometrical sentence is valid, but we do not, and probably shall never know, any such method for deciding whether a sentence of this sort is elementarily provable.*

The differences between \mathcal{E}_2 and \mathcal{E}_2'' vanish when we restrict ourselves to universal sentences. In fact, we have

THEOREM 7. *A universal sentence formulated in \mathcal{E}_2 is valid in \mathcal{E}_2 if and only if it is valid in \mathcal{E}_2'' .*

To prove this we recall that every ordered field can be extended to a real closed field. Hence, by Theorems 1 and 6, every model of \mathcal{E}_2'' can be extended to a model of \mathcal{E}_2 . Consequently, every universal sentence which is valid in \mathcal{E}_2 is also valid in \mathcal{E}_2'' ; the converse is obvious. (An even simpler proof of Theorem 7, and in fact a proof independent of Theorem 1, can be based upon the lemma by which every finite subsystem of an ordered field can be isomorphically embedded in the ordered field of real numbers.)

Theorem 7 remains valid if we remove A13' from the axiom system of \mathcal{E}_2'' (and it applies even to some still weaker axiom systems). Thus we see that every elementary universal sentence which is valid in \mathcal{E}_2 can be proved without any help of the continuity axioms. The result extends to

all the sentences which may not be universal when formulated in \mathcal{E}_2 but which, roughly speaking, become universal when expressed in the notation of Cartesian spaces $\mathcal{C}_2(\mathfrak{F})$.

As an immediate consequence of Theorems 3 and 7 we obtain:

THEOREM 8. *The theory \mathcal{E}_2'' is decidable with respect to the set of its universal sentences.*

This means that there is a mechanical method for deciding in each particular case whether or not a given universal sentence formulated in the theory \mathcal{E}_2'' holds in every model of this theory.

We could discuss some further theories related to \mathcal{E}_2 , \mathcal{E}_2' , and \mathcal{E}_2'' ; e.g., the theory \mathcal{E}_2''' which has the same symbolism as \mathcal{E}_2' and the same axiom system as \mathcal{E}_2'' . The problem of deciding which of the various formal conceptions of elementary geometry is closer to the historical tradition and the colloquial usage of this notion seems to be rather hopeless and deprived of broader interest. The author feels that, among these various conceptions, the one embodied in \mathcal{E}_2 distinguishes itself by the simplicity and clarity of underlying intuitions and by the harmony and power of its metamathematical implications.

Bibliography

- [1] HILBERT, D., *Grundlagen der Geometrie*. Eighth edition, with revisions and supplements by P. BERNAYS, Stuttgart 1956, III+251 pp.
- [2] ROBINSON, J., *Definability and decision problems in arithmetic*. Journal of Symbolic Logic, vol. 14 (1949), pp. 98–114.
- [3] ROBINSON, R. M., *Binary relations as primitive notions in elementary geometry*. This volume, pp. 68–85.
- [4] SCHWABHÄUSER, W., *Über die Vollständigkeit der elementaren euklidischen Geometrie*. Zeitschrift für mathematische Logik und Grundlagen der Mathematik, vol. 2 (1956), pp. 137–165.
- [5] SCOTT, D., *Dimension in elementary Euclidean geometry*. This volume, pp. 53–67.
- [6] SZMIELEW, W., *Some metamathematical problems concerning elementary hyperbolic geometry*. This volume, pp. 30–52.
- [7] TARSKI, A., *A decision method for elementary algebra and geometry*. Second edition, Berkeley and Los Angeles 1951, VI+63 pp.
- [8] ———, *Contributions to the theory of models*. Indagationes Mathematicae, vol. 16 (1954), pp. 572–588, and vol. 17 (1955), pp. 56–64.

- [9] ——— MOSTOWSKI, A., and ROBINSON, R. M., *Undecidable theories*. Amsterdam 1953, XI+98 pp.
- [10] ——— and VAUGHT, R. L., *Arithmetical extensions of relational systems*. *Compositio Mathematica*, vol. 13 (1957), pp. 81–102.
- [11] VAN DER WAERDEN, B. L., *Modern Algebra*. Revised English edition, New York, 1953, vol. 1, XII+264 pp.

SOME METAMATHEMATICAL PROBLEMS CONCERNING ELEMENTARY HYPERBOLIC GEOMETRY

WANDA SZMIELEW

*University of Warsaw, Warsaw, Poland, and Institute for Basic Research in Science,
University of California, Berkeley, California, U.S.A.*

Introduction. In this paper we shall be concerned with a formalized system \mathcal{H}_n of elementary n -dimensional hyperbolic (Bolyai-Lobachevskian) geometry. Throughout the paper we shall use the notation introduced by Tarski in [5]. In particular the system \mathcal{H}_n has the same logical structure and the same symbolism as Tarski's system \mathcal{E}_2 of elementary Euclidean geometry. In case $n = 2$ it differs from \mathcal{E}_2 only in that Euclid's axiom, A8, has been replaced by its negation; for $n > 2$ the dimension axioms, A11 and A12, should, in addition, be appropriately modified. The aim of this paper is to extend to the system \mathcal{H}_n the fundamental metamathematical results stated in [5] for the system \mathcal{E}_2 .¹

The paper is divided into three sections. In Section 1 we shall indicate how the solutions of the metamathematical problems in which we are interested can be obtained by means of a familiar algorithm, the end-calculus of Hilbert (cf. [1], pp. 159ff.). In Section 2 we shall construct a new geometrical algorithm, the hyperbolic calculus of segments, which will prove to provide a convenient apparatus for a new solution of the same problems. The results established in Sections 1 and 2 have interesting implications for some related geometrical systems, in fact, for elementary absolute geometry (i.e., the common part of elementary Euclidean and hyperbolic geometries) and for non-elementary hyperbolic geometry. These implications will be discussed in Section 3.²

1. Hilbert-Szász Spaces. In [1] Hilbert gives an outline of his end-calculus, defines in its terms the coordinates of straight lines and points

¹ The results of this paper were obtained while the author was working in the University of California, Berkeley on a research project in the foundations of mathematics sponsored by the U.S. National Science Foundation.

² All the observations which will be given in Section 1 and those concerning absolute geometry in Section 3 have been made jointly by Tarski and the author.

and establishes an analytic condition for a point to lie on a straight line. The whole discussion is done in a system of hyperbolic geometry included in \mathcal{H}_2 . In [3] Szász somewhat modifies Hilbert's construction and moreover establishes an analytic formula for the distance between two points. The latter formula is essential for our purposes and therefore we shall refer in what follows to [3] and not to [1].

The discussion in [3] leads to an important class of models of \mathcal{H}_2 which can be obtained by means of the following algebraic construction: Consider an arbitrary ordered field $\mathfrak{F} = \langle F, +, \cdot, \leq \rangle$. For any ordered triples $x = \langle x_1, x_2, x_3 \rangle$ and $y = \langle y_1, y_2, y_3 \rangle$ in $F \times F \times F$ let

$$\Phi(x, y) = x_1 \cdot y_1 + x_2 \cdot y_2 - x_3 \cdot y_3.$$

By $A_{\mathfrak{F}}$ we denote the subset of $F \times F \times F$ consisting of all those triples x , for which

$$\Phi(x, x) = -1 \text{ and } x_3 > 0.$$

By $B_{\mathfrak{F}}$ (the *betweenness relation*) we denote the ternary relation which holds among the triples x, y, z in $A_{\mathfrak{F}}$ if and only if

$$\Phi(u, u) > 0, \Phi(u, x) = 0, \Phi(u, y) = 0, \Phi(u, z) = 0 \text{ for some } u \in F \times F \times F$$

and moreover

$$\Phi(x, y) \geq \Phi(x, z) \text{ and } \Phi(y, z) \geq \Phi(x, z).$$

Finally, by $D_{\mathfrak{F}}$ (the *equidistance relation*) we denote the quaternary relation which holds among the triples x, y, z, u in $A_{\mathfrak{F}}$ if and only if

$$\Phi(x, y) = \Phi(z, u).$$

The system $\langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ thus obtained will be denoted by $\mathfrak{H}_2(\mathfrak{F})$ and will be referred to as the *two-dimensional Hilbert-Szász space over the field \mathfrak{F}* .

As a direct consequence of Szász' discussion the following result is obtained: Every model of \mathcal{H}_2 is isomorphic with the space $\mathfrak{H}_2(\mathfrak{F})$ over some Euclidean field \mathfrak{F} . By supplementing the argument of Szász we easily show, by means of the elementary continuity axioms, A13 (see [5], p. 20), that the field \mathfrak{F} is real closed. Since \mathcal{H}_2 has a model, then for some real closed field \mathfrak{F} the space $\mathfrak{H}_2(\mathfrak{F})$ is a model of \mathcal{H}_2 . And since, by a fundamental result of Tarski in [4], any two real closed fields \mathfrak{F}' and \mathfrak{F}'' are elementarily equivalent, so are also spaces $\mathfrak{H}_2(\mathfrak{F}')$ and $\mathfrak{H}_2(\mathfrak{F}'')$, and

consequently each of the spaces $\mathfrak{H}_2(\mathfrak{F})$ is a model of \mathcal{H}_2 . This clearly applies to all the systems isomorphic with $\mathfrak{H}_2(\mathfrak{F})$ as well. Thus we have arrived at the following

THEOREM 1.1. (REPRESENTATION THEOREM). *A system $\mathfrak{M} = \langle A, B, D \rangle$ is a model of \mathcal{H}_2 if and only if it is isomorphic with the Hilbert-Szász space $\mathfrak{H}_2(\mathfrak{F}) = \langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ over some real closed field \mathfrak{F} .*

Theorem 1.1 implies as a corollary

THEOREM 1.2. *The theory \mathcal{H}_2 is complete and decidable but not finitely axiomatizable.*

The proof of Theorem 1.2 is quite analogous to the proof of the corresponding results (Theorems 2, 3, 4) for the system \mathcal{E}_2 in [5].

In this way we have established the fundamental metamathematical properties of two-dimensional elementary hyperbolic geometry. The extension of these results to n -dimensional geometries does not seem to present any essential difficulty.

2. Klein Spaces and Hyperbolic Calculus of Segments. In this section we wish to establish fundamental metamathematical properties of \mathcal{H}_n by using in the representation theorem, instead of the Hilbert-Szász models, the much more familiar and intuitively simpler *Klein models*. We could try to derive the new representation theorem from the old one by showing in a purely algebraic way that every Klein model is isomorphic with some Hilbert-Szász model, and conversely. We prefer, however, to obtain this result by means of a direct procedure, and to this end we construct a special geometrical algorithm, which will be called the *hyperbolic calculus of free segments*. This algorithm seems to present some geometrical interest independent of any metamathematical applications and to be conceptually simpler than the end-calculus of Hilbert.

Consider a model $\mathfrak{M} = \langle A, B, D \rangle$ of \mathcal{H}_n ($n \geq 2$) formed by an arbitrary set A , a ternary relation B (the betweenness relation), and a quaternary relation D (the equidistance relation) among elements (points) of A . By a *segment* we understand any non-ordered couple pq of two distinct points p, q in A . Two segments pq and rs are *congruent* (in symbols, $pq \cong rs$) if and only if $D(pqrs)$. The set of all segments congruent to a given segment pq is called the *free segment* determined by pq and is denoted by $[pq]$. Free segments will be represented by variables X, Y, Z, \dots and the set of all free segments will be denoted by \mathcal{S} . We wish to define a binary relation

\leq between elements of \mathcal{S} and two binary operations $+$ and \cdot on elements of \mathcal{S} in such a way that the rectangular coordinates introduced on the base of the resulting calculus function as the Beltrami coordinates, and, in fact, lead to Klein model.

To obtain appropriate definitions let us assume for a while that \mathfrak{M} is a model, not only of \mathcal{H}_2 , but of full two-dimensional hyperbolic geometry with the non-elementary axiom of continuity (e.g. the ordinary Klein model). As is well known, in such a model \mathfrak{M} we can correlate with every angle PQ a real number $\mu(PQ)$, $0 < \mu(PQ) < \pi$, called the *measure* of PQ . The angle PQ is understood here as the non-ordered pair of half-lines P and Q which are supposed to be non-collinear and to have a common origin. Hence we can define in \mathfrak{M} the *Lobachevskian function* Π , which assigns a real number $\Pi(X)$, $0 < \Pi(X) < \frac{\pi}{2}$, to every free segment X .

In fact, given an oriented straight line L (Figure 1), a point p not on L ,

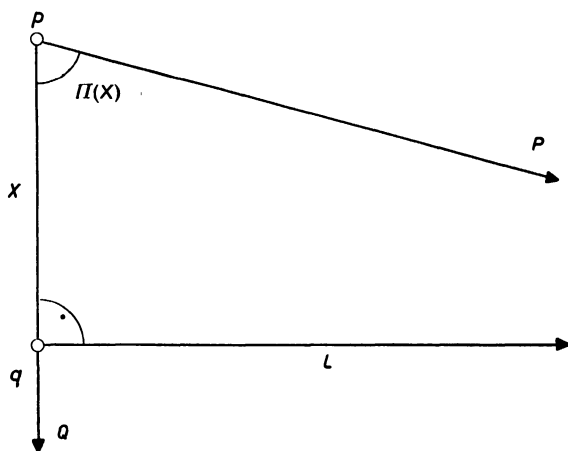


Fig. 1

the perpendicular projection q of p upon L , and the half-line P with origin p and parallel to L , if Q is the half-line pq and $X = [pq]$, then $\Pi(X) = \mu(PQ)$. The *Beltrami coordinates* of an arbitrary point p of the model \mathfrak{M} , if different from 0, are numbers of the form $\pm \cos \Pi(X)$, $\pm \cos \Pi(Y)$, where X and Y are two free segments correlated with point p . Using this fact we define the relation \leq and the operations $+$

and \cdot for elements of \mathcal{S} by the following conditions

- (I) $X \leq Y$ if and only if $\cos \Pi(X) \leq \cos \Pi(Y)$,
- (II) $X + Y = Z$ if and only if $\cos \Pi(X) + \cos \Pi(Y) = \cos \Pi(Z)$,
- (III) $X \cdot Y = Z$ if and only if $\cos \Pi(X) \cdot \cos \Pi(Y) = \cos \Pi(Z)$.

We shall show that these definitions can be replaced by equivalent ones formulated entirely in terms of the relations B and D . This will make it possible to extend the definitions to an arbitrary model \mathfrak{M} .

Relation \leq . In view of (I) (since both functions, \cos in the interval $(0, \frac{\pi}{2})$ and Π , are decreasing) \leq is the ordinary *less than or equal to* relation; speaking precisely

- (I') $X \leq Y$ if and only if $B(pqr)$, $[pq] = X$, and $[pr] = Y$, for some $p, q, r \in A$.

As usual the symbol \geq will denote the relation converse to \leq .

In defining the operations $+$ and \cdot , and in deducing their fundamental properties we shall use the notions of a proper or improper right triangle and of a proper or improper right quadrangle (i.e., a quadrangle with three right angles). For our purposes it is convenient to introduce these notions in the following way:

Given three non-collinear points p, q, r , we say that the ordered triple pqr is a (*proper*) *right triangle* if and only if $\sphericalangle pqr$ is a right angle.

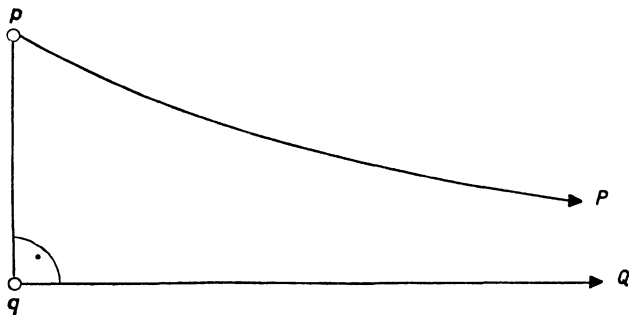


Fig. 2

Given two distinct points p and q , a half-line P with origin p , and a half-line Q with origin q (Figure 2), we say that the ordered quadruple

$PpqQ$ is an *improper right triangle* if and only if the half-lines qp and Q form a right angle and $P \parallel Q$. It is clear that points p and q uniquely determine half-lines P and Q .

Given four points p, q, r, s , no three of which are collinear, we say that the ordered quadruple $pqrs$ is a (*proper*) *right quadrangle* if and only if $\sphericalangle spq$, $\sphericalangle pqr$, and $\sphericalangle qrs$ are three right angles. It is clear that there are non-collinear points p, q, r , such that $\sphericalangle pqr$ is a right angle and for which there is no point s such that $pqrs$ is a right quadrangle.

Given three distinct points p, q, r , a half-line P with origin p , and a half-line R with origin r (Figure 3), we say that the ordered quintuple $PpqrR$ is an *improper right quadrangle* if and only if half-lines P and pq , qp and qr , rq and R form three right angles and $P \parallel R$. It is clear that points p and q determine uniquely the half-line P , the point r , and the half-line R .

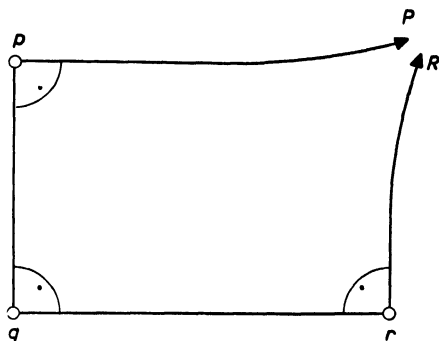


Fig. 3

Before defining the operations \oplus and \odot in terms of the relations B and D we first introduce four auxiliary operations on elements of \mathcal{S} , in fact, two binary operations, \oplus and \odot , and two unary operations, \mathbf{R} and \mathbf{C} .

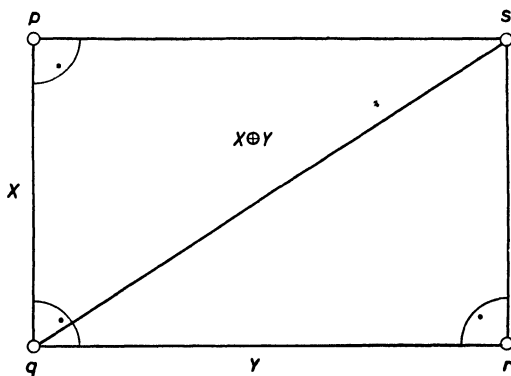


Fig. 4

Operation \oplus . Given two free segments X and Y , consider the free segment Z constructed in the following way: For some right quadrangle $pqrs$, let $X = [pq]$, $Y = [qr]$, and $Z = [qs]$ (Figure 4). Clearly, the segment Z thus defined not always exists (since the right quadrangle $pqrs$ not always can be constructed). If however Z exists, it is uniquely determined by X and Y (independent of the choice of $pqrs$) and we then put $X \oplus Y = Z$. To express the fact that $X \oplus Y$ does, or does not exist, we shall respectively write $X \oplus Y \in \mathcal{S}$, $X \oplus Y \notin \mathcal{S}$.

Operation \odot .³ Given two free segments X and Y , we consider a right triangle pqr with $X = [pq]$ and $Y = [qr]$ (Figure 5), and we put $X \odot Y =$

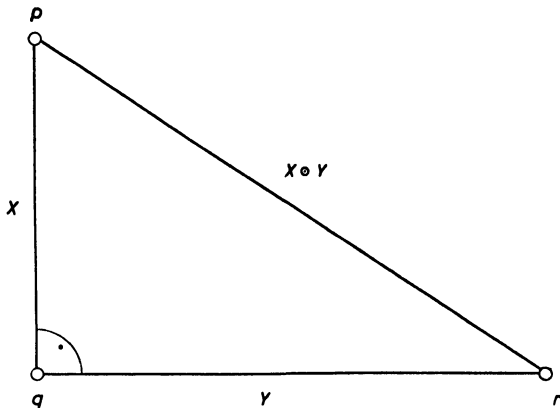


Fig. 5

$[pr]$. The operation \odot thus defined is always performable, i.e., we have $X \odot Y \in \mathcal{S}$ for any $X, Y \in \mathcal{S}$.

It is worth while to notice that both the operations \oplus and \odot have sense in absolute geometry and that they coincide in Euclidean geometry.

Operation \mathbf{R} . Given a free segment X , we consider an isocles right triangle pqr with $X = [pr]$ (Figure 6), and we put $\mathbf{R}X = [pq] = [qr]$. Clearly the operation \mathbf{R} is always performable. $\mathbf{R}X$ can be referred to as the *square root* of X .

Operation \mathbf{C} . Given a free segment X , we consider an improper right quadrangle $PpqrR$ with $X = [pq]$ (Figure 7), and we put $\mathbf{C}X = [qr]$.

³ This operation was studied by Hjelmlev in [2].

Obviously the operation C is always performable. CX can be referred to as the *complement* of X .

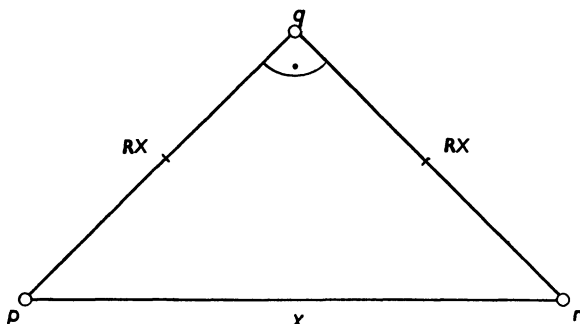


Fig. 6

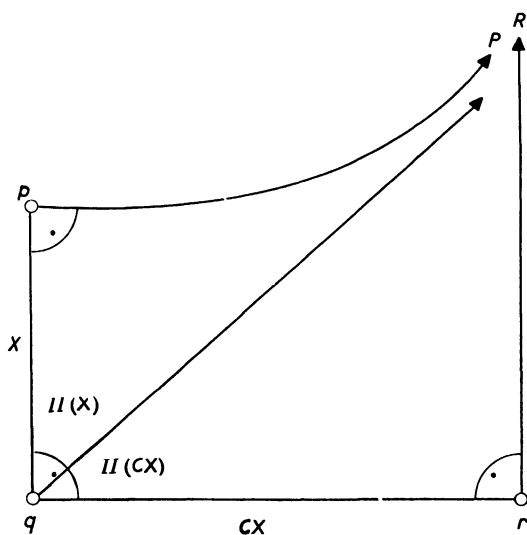


Fig. 7

Clearly, the four operations just defined can be characterized in terms of the primitive relations B and D .

Using some well known theorems of hyperbolic geometry we can

easily establish in \mathfrak{M} the formulas

$$\cos^2 \Pi(X) + \cos^2 \Pi(Y) = \cos^2 \Pi(X \oplus Y),$$

$$\sin \Pi(X) \cdot \sin \Pi(Y) = \sin \Pi(X \odot Y),$$

$$\sin \Pi(X) = \sin^2 \Pi(RX),$$

$$\Pi(X) + \Pi(CX) = \frac{\pi}{2}.$$

By definitions (II) and (III) these formulas imply at once the following equivalences:

$$(II') \quad X + Y = Z \quad \text{if and only if} \quad CRCX \oplus CRCY = CRCZ,$$

$$(III') \quad X \cdot Y = Z \quad \text{if and only if} \quad CX \odot CY = CZ.$$

We now return to the original model \mathfrak{M} of \mathcal{H}_n . In this model we introduce the auxiliary operations \oplus , \odot , R , C (in the definitions of \oplus and R it should be additionally mentioned that $pqrs$ and $PpqrR$ are quadrangles on a plane) and assume equivalences (I'), (II'), (III') as definitions of \leq , $+$, \cdot .

We shall now establish the fundamental properties of the system $\mathfrak{S} = \langle S, +, \cdot, \leq \rangle$. A detailed discussion will be given only for the case $n \geq 3$, thus using (when needed) three-dimensional constructions. Some remarks concerning the case $n = 2$ will be given later.

In Lemma 2.1 we state some fundamental properties of the relation \leq and the auxiliary operations.

LEMMA 2.1. *The system $\langle S, \oplus, \odot, R, C, \leq \rangle$ satisfies the following conditions:*

- (i) $\langle S, \leq \rangle$ is a non-empty simply ordered system;
- (ii) if $X, Y, X \oplus Y \in S$, then $X \oplus Y = Y \oplus X$;
- (iii) if $X, Y, Z, X \oplus Y, (X \oplus Y) \oplus Z \in S$, then $Y \oplus Z \in S$ and $(X \oplus Y) \oplus Z = X \oplus (Y \oplus Z)$;
- (iv) if $X, Z \in S$, then $X \leq Z$ if and only if $X = Z$ or else $X \oplus Y = Z$ for some $Y \in S$;
- (v) if $X, Y \in S$, then $X \odot Y \in S$ and $X \odot Y = Y \odot X$;
- (vi) if $X, Y, Z \in S$, then $(X \odot Y) \odot Z = X \odot (Y \odot Z)$;

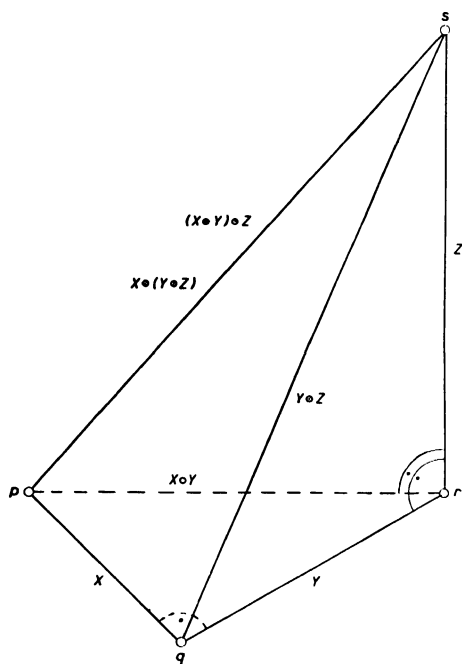


Fig. 9

the straight line L which passes through the point r , is perpendicular to the straight line pr , and lies in the plane prs . Then $[pt] = X \oplus Y$ and $[pu] = (X \oplus Y) \oplus Z$. Furthermore $qpwu$ proves to be a right quadrangle; using this fact $rpsw$ is shown to be a right quadrangle as well. Hence $[pw] = Y \oplus Z$ and $[pu] = X \oplus (Y \oplus Z)$. Consequently $(X \oplus Y) \oplus Z = X \oplus (Y \oplus Z)$, what was to be proved.

To derive Postulate (vi) (the associative law for \odot) let $X, Y, Z \in \mathcal{S}$ and let p, q, r, s be four distinct points (Figure 9) satisfying the conditions: (δ) $[pq] = X$, $[qr] = Y$, $[rs] = Z$, and (ϵ) $\nless pqr$, $\nless prs$, $\nless qrs$ are three right angles. Then

$[pr] = X \odot Y$, $[qs] = Y \odot Z$, $[ps] = (X \odot Y) \odot Z$, and $\nless pqs$ is a right angle. Hence $[ps] = X \odot (Y \odot Z)$ and consequently $(X \odot Y) \odot Z = X \odot (Y \odot Z)$.⁴ The proof of Lemma 2.1 has thus been completed.

By the next two Lemmas the discussion of the properties of the operations \cdot and $+$ reduces to that of the properties of the operations \odot and \oplus respectively.

LEMMA 2.2. *The function C maps the system $\langle \mathcal{S}, \cdot, \leq \rangle$ isomorphically onto the system $\langle \mathcal{S}, \odot, \geq \rangle$.*

This lemma follows directly from Lemma 2.1(x)(xi) and the definition of \cdot .

LEMMA 2.3. *The function CRC maps the system $\langle \mathcal{S}, +, \cdot, \leq \rangle$ isomorphically onto the system $\langle \mathcal{S}, \oplus, \cdot, \leq \rangle$.*

⁴ The argument used in the proof of (vi) can be found in [2], p. 5.

PROOF. By Lemmas 2.2, 2.1(viii)(ix) and the definition of $+$, the function \mathbf{CRC} maps the system $\langle \mathcal{S}, +, \leq \rangle$ isomorphically onto the system $\langle \mathcal{S}, \oplus, \leq \rangle$. To complete the proof it is sufficient to show that

$$(1) \quad X \cdot Y = Z \quad \text{if and only if} \quad \mathbf{CRC}X \cdot \mathbf{CRC}Y = \mathbf{CRC}Z.$$

From Lemma 2.1(v)–(viii) we easily derive the formula

$$\mathbf{R}(X \odot Y) = \mathbf{R}X \odot \mathbf{R}Y,$$

which, together with the definition of \cdot and Lemma 2.1(x), gives us the required equivalence (1).

The next lemma provides a new geometrical construction by means of which the operation \cdot can be obtained. This lemma will eventually lead to the distributive law for \cdot under $+$; it also will be helpful in setting up the foundations of the theory of proportion (see Lemma 2.11).

LEMMA 2.4. Let $PpqQ$ be an improper right triangle. Furthermore, let $r \in P$ and let s be the perpendicular projection of r upon the straight line pq . Under these assumptions, if $[pq] = X$ and $[pr] = Y$, then $[ps] = X \cdot Y$ (Figure 10).

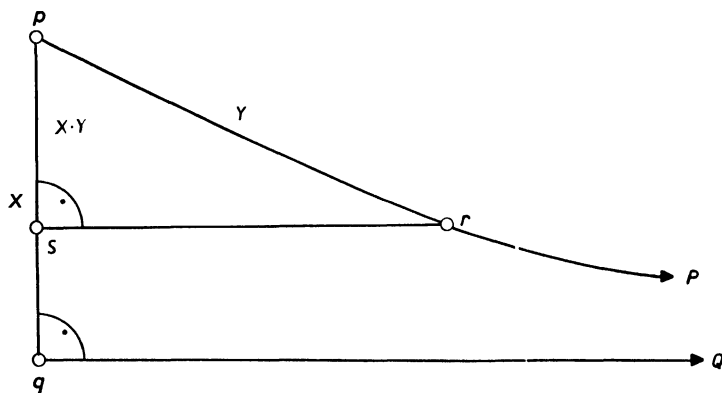


Fig. 10

PROOF. We assume that $[pq] = X$ and $[pr] = Y$. Consider four points t, p_1, r_1, s_1 and three half-lines T, R_1, S_1 (Figure 11) which satisfy the following conditions: (α) $p_1 \neq p$, the straight line pp_1 is perpendicular to the plane pqr , and $[pp_1] = CY$, and (β) $Ppp_1r_1R_1, Qqp_1tT$, and $Ttp_1s_1S_1$

PROOF. Let $X, U \in \mathcal{S}$. We pick an improper right quadrangle $QqPq_1Q_1$ for which $[pq] = X$ (Figure 12). Then $[pq_1] = CX$. On the half-line P with origin p and parallel to the half-lines Q and Q_1 we choose a point r in such a way that $[pr] = U$. Let s and s_1 be perpendicular projections of r upon the straight lines pq and pq_1 . Then $sp s_1 r$ is a right quadrangle, and

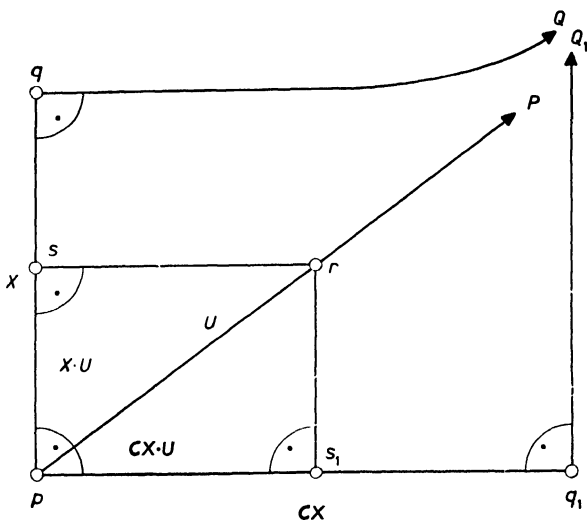


Fig. 12

by Lemma 2.4, we have $[ps] = X \cdot U$ and $[ps_1] = CX \cdot U$. Hence $X \cdot U \oplus X \cdot CU = U$, which completes the proof.

As an immediate consequence of Lemmas 2.1(i)–(vii), 2.2, 2.3, 2.1(xii)–(xiii), and 2.5 we obtain the fundamental theorem on the calculus of free segments.

THEOREM 2.6. *For every model \mathfrak{M} of \mathcal{H}_n ($n \geq 3$), the system $\mathfrak{S} = \langle \mathcal{S}, +, \cdot, \leq \rangle$ satisfies the following conditions:*

- (i) $\langle \mathcal{S}, \leq \rangle$ is a non-empty simply ordered system;
- (ii) if $X, Y, X + Y \in \mathcal{S}$, then $X + Y = Y + X$;
- (iii) if $X, Y, Z, X + Y, (X + Y) + Z \in \mathcal{S}$, then $Y + Z \in \mathcal{S}$ and $(X + Y) + Z = X + (Y + Z)$;

- (iv) if $X, Z \in \mathcal{S}$, then $X \leq Z$ if and only if $X = Z$ or else $X + Y = Z$ for some $Y \in \mathcal{S}$;
- (v) if $X, Y \in \mathcal{S}$, then $X \cdot Y \in \mathcal{S}$ and $X \cdot Y = Y \cdot X$;
- (vi) if $X, Y, Z \in \mathcal{S}$, then $(X \cdot Y) \cdot Z = X \cdot (Y \cdot Z)$;
- (vii) if $X, Z \in \mathcal{S}$, then $Z \leq X$ if and only if $Z = X$ or else $X \cdot Y = Z$ for some $Y \in \mathcal{S}$.
- (viii) if $X \in \mathcal{S}$, then there is a $Y \in \mathcal{S}$ such that: (α) $X + Y \notin \mathcal{S}$, (β) if $Z \in \mathcal{S}$ and $Z \leq Y$, then $X + Z \in \mathcal{S}$, and (γ) $X \cdot U + Y \cdot U = U$ for every $U \in \mathcal{S}$.

NOTE 2.7. Theorem 2.6 can be extended to the case $n = 2$.

In fact, Lemmas 2.1(iii), 2.1(vi), and 2.4 are the only ones in proofs of which three dimensional constructions are involved. These constructions should now be replaced by two-dimensional ones. Unfortunately, a direct two-dimensional proof of Lemma 2.4 is still lacking.⁵ We know, however, an indirect two-dimensional proof of this lemma; it is based upon one of the fundamental results of Section 1, namely the completeness of \mathcal{H}_2 (see Theorem 1.2). On the other hand, we know two direct two-dimensional arguments which lead from Lemma 2.4 to Lemmas 2.1(iii) and 2.1(vi), respectively. As opposed to the three-dimensional proofs of Lemmas 2.1(iii) and 2.1(vi) which have a quite elementary character, these two-dimensional arguments are rather involved and refer to deep properties of the plane. Lack of space prevents us from outlining these constructions.

As a consequence of Theorem 2.6, Note 2.7 and the elementary continuity axioms, we obtain by purely algebraic argument

THEOREM 2.8. *For every model \mathfrak{M} of \mathcal{H}_n ($n \geq 2$), the system $\mathfrak{S} = \langle \mathcal{S}, +, \cdot, \leq \rangle$ can be imbedded in a real closed field $\mathfrak{F} = \langle \mathbf{F}, +, \cdot, \leq \rangle$ in such a way that \mathcal{S} consists of all those elements $X \in \mathbf{F}$ for which $0 < X < 1$ (where 0 is the zero element and 1 is the unit element of the field \mathfrak{F}). In fact, \mathfrak{F} is up to isomorphism uniquely determined by \mathfrak{S} .*

The proof of this theorem is easy, though lengthy and laborious. Postulate (viii) (of Theorem 2.6) plays an essential role in showing that \mathcal{S} is the set of all elements of \mathbf{F} between 0 and 1. While the last part of (viii)

⁵ See Footnote 6 on page 51.

is a particular case of the distributive law, (viii) plays also an essential role in the derivation of this law in its general form.

From now on we assume that the field \mathfrak{F} involved in Theorem 2.8 has been fixed and we apply to it the familiar field-theoretical notation. In particular, the operations $+$ and \cdot are now understood to be performable on arbitrary elements of the field and not only on free segments.

Theorem 2.8 essentially completes our outline of the calculus of free segments. We shall need however a few further lemmas of a related character before we turn, in Theorems 2.15 and 2.16, to the metamathematical discussion of systems \mathcal{H}_n .

LEMMA 2.9. *If $X, Y, Z \in \mathcal{S}$, then*

- (i) $X \odot Y = Z$ *if and only if* $CX \cdot CY = CZ$;
- (ii) $X \oplus Y = Z$ *if and only if* $X^2 + Y^2 = Z^2$;
- (iii) $CX = \sqrt{1 - X^2}$.

PROOF. The equivalence (i) is an immediate consequence of the definition of \odot and Lemma 2.1(x).

By (i) and Lemma 2.1(viii)–(x) we get

$$(2) \quad \mathbf{CRCX} \cdot \mathbf{CRCX} = X.$$

From the definition of $+$ and formulas (1) (on page 41) and (2) we easily derive the equivalence (ii).

The formula (iii) follows at once from (ii) and Lemma 2.5.

LEMMA 2.10. *Let pqr be a right quadrangle and let $X = [pq]$, $Y = [qr]$, $Z = [rs]$. Then we have $X = CY \cdot Z$.*

PROOF. Let $U = [qs]$ (Figure 13). Then, in agreement with the definitions of \oplus and \odot , and by Lemma 2.9, we have

$$U^2 = X^2 + Y^2 \text{ and } 1 - U^2 = (1 - Y^2) \cdot (1 - Z^2).$$

Comparing these two formulas and applying Lemma 2.9 (iii) we obtain the conclusion.

LEMMA 2.11. *For $i = 1, 2$, let $p_i r_i s_i$ be right triangles and let $Y_i = [p_i r_i]$ and $Z_i = [p_i s_i]$. Under these assumptions, if the angles at the vertices p_1 and p_2 are congruent, then $Y_1 \cdot Z_2 = Y_2 \cdot Z_1$.*

PROOF. Assume that $\sphericalangle r_1 p_1 s_1 \cong \sphericalangle r_2 p_2 s_2$ and let P_i be the half-line $p_i r_i$ (Figure 14). The triangle $p_i r_i s_i$ determines uniquely a point q_i on the

half-line p_1s_1 and a half-line Q_1 with origin q_1 such that $P_1p_1q_1Q_1$ is an improper right triangle. Then $[p_1q_1] = [p_2q_2]$. Putting $[p_1q_1] = [p_2q_2] = X$ and applying Lemma 2.4 we get the formula

$$Y_1 \cdot Z_2 = Y_1 \cdot (X \cdot Y_2) = Y_2 \cdot (X \cdot Y_1) = Y_2 \cdot Z_1,$$

which completes the proof.

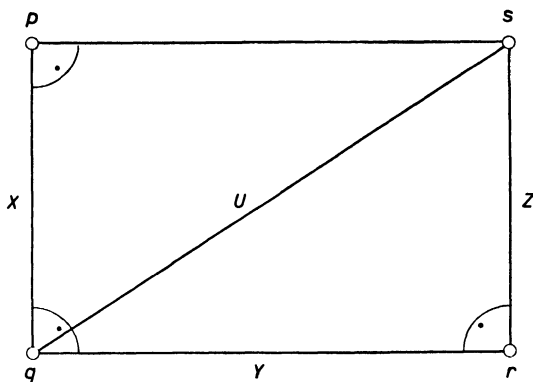


Fig. 13

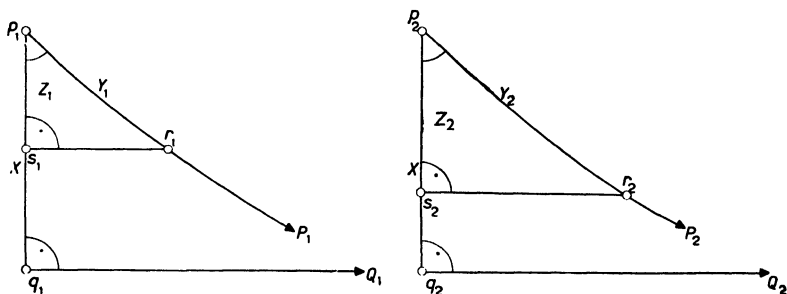


Fig. 14

LEMMA 2.12. Let pqr be a right triangle, let s be the perpendicular projections of q upon the straight line pr , and let $X = [ps]$, $Y = [rs]$, $Z = [pr]$, $U = [pq]$, and $V = [qr]$. Then we have:

- (i) $X \cdot Z = U \cdot U$,
- (ii) $X \odot Z = Y \odot U \odot U$

(Figure 15).

PROOF. Formula (i) follows immediately from Lemma 2.11. Let $W = [qs]$. Then

$$X \odot V = X \odot (Y \odot W) = Y \odot (X \odot W) = Y \odot U,$$

and consequently

$$X \odot Z = X \odot (U \odot V) = (X \odot V) \odot U = Y \odot U \odot U;$$

thus we arrive at formula (ii).

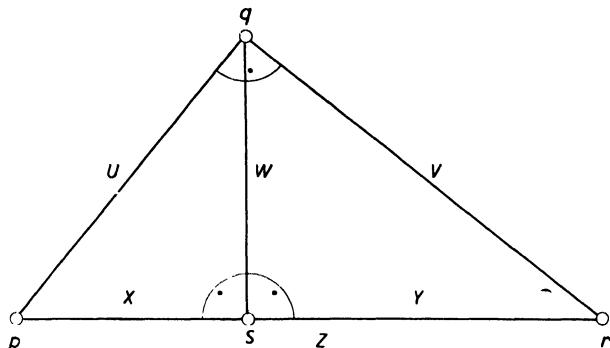


Fig. 15

LEMMA 2.13. *Given three distinct points p, s, r for which $B(psr)$, let $X = [ps]$, $Y = [sr]$, and $Z = [pr]$. We then have*

$$(i) \quad CY = \frac{CX \cdot CZ}{1 - X \cdot Z}$$

and

$$(ii) \quad CZ = \frac{CX \cdot CY}{1 + X \cdot Y}.$$

PROOF. To derive (i) we take a point q in such a way that pqr is a right triangle and s is the perpendicular projection of q upon the straight line pr (Figure 15). Let $U = [pq]$. Then by Lemma 2.12(i), we have $X \cdot Z = U^2$, i.e.,

$$(3) \quad 1 - X \cdot Z = (CU)^2,$$

and, by Lemma 2.12(ii), we get $X \odot Z = Y \odot U \odot U$, which by Lemma

2.9(i) implies

$$(4) \quad CX \cdot CZ = CY \cdot (CU)^2.$$

From (3) and (4) we obtain at once the desired formula.

From (i) and the inequality $X < Z$ (which obviously follows from the hypothesis) we derive (ii) by means of a simple algebraic transformation.

LEMMA 2.14. *Given four distinct points p, q, r, s , we have*

- (i) $B(pqr)$ if and only if $C[pr] = \frac{C[pq] \cdot C[qr]}{1 + [pq] \cdot [qr]}$,
- (ii) $D(pqrs)$ if and only if $C[pq] = C[rs]$.

PROOF. Formula (i) follows Lemma 2.13, in one direction directly, in the other direction by a simple argument. Formula (ii) is obvious.

The metamathematical discussion begins with the representation theorem.

Let $\mathfrak{F} = \langle F, +, \cdot, \leq \rangle$ be an arbitrary ordered field. By the *n-dimensional Klein space* $\mathfrak{K}_n(\mathfrak{F})$ over the field \mathfrak{F} we understand the system $\langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ constructed in the following way: $A_{\mathfrak{F}}$ is the set of all ordered n -tuples $x = \langle x_1, x_2, \dots, x_n \rangle$ in $F \times F \times \dots \times F$ (n times) for which

$$x_1^2 + x_2^2 + \dots + x_n^2 < 1.$$

For any ordered n -tuples $x = \langle x_1, x_2, \dots, x_n \rangle$ and $y = \langle y_1, y_2, \dots, y_n \rangle$ in $A_{\mathfrak{F}}$ let

$$x \cdot y = \sum_{i=1}^n x_i \cdot y_i$$

(thus $-1 < x \cdot y < 1$) and

$$\Psi(x, y) = \frac{(1 - x \cdot x) \cdot (1 - y \cdot y)}{(1 - x \cdot y)^2}.$$

We always have $\Psi(x, y) \leq 1$. The betweenness relation $B_{\mathfrak{F}}$ among any three n -tuples x, y, z in $A_{\mathfrak{F}}$ is characterized by the formula

$$\Psi(x, z) = \frac{\Psi(x, y) \cdot \Psi(y, z)}{(1 + \sqrt{1 - \Psi(x, y)} \cdot \sqrt{1 - \Psi(y, z)})^2}.$$

The equidistance relation $D_{\mathfrak{F}}$ among any four n -tuples x, y, z, u in $A_{\mathfrak{F}}$ is

characterized by the formula

$$\Psi(x, y) = \Psi(z, u).$$

THEOREM 2.15. (REPRESENTATION THEOREM). *A system $\mathfrak{M} = \langle A, B, D \rangle$ is a model of \mathcal{H}_n ($n \geq 2$) if and only if it is isomorphic with the Klein space $\mathfrak{K}_n(\mathfrak{F}) = \langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ over some real closed field \mathfrak{F} .*

PROOF. It is well known that the Klein space $\mathfrak{K}_n(\mathfrak{R})$ over the ordered field \mathfrak{R} of real numbers is a model for \mathcal{H}_n . Hence, by the result of Tarski used in the proof of Theorem 1.1, the same applies to all the spaces $\mathfrak{K}_n(\mathfrak{F})$ where \mathfrak{F} is a real closed field, as well as to all isomorphic systems.

To prove the theorem in the opposite direction consider an arbitrary model $\mathfrak{M} = \langle A, B, D \rangle$ of \mathcal{H}_n and the correlated system $\mathfrak{S} = \langle S, +, \cdot, \leq \rangle$. By Theorem 2.12, the system \mathfrak{S} can be imbedded in a real closed field $\mathfrak{F} = \langle F, +, \cdot, \leq \rangle$, and we can construct the corresponding Klein model $\mathfrak{K}_n(\mathfrak{F}) = \langle A_{\mathfrak{F}}, B_{\mathfrak{F}}, D_{\mathfrak{F}} \rangle$ over the field \mathfrak{F} . We introduce in this model a rectangular coordinate system (each of the n coordinates of a point p being of the form $\pm U$ where $U \in S$). It is easy to check, that by correlating with every point p of A the ordered n -tuple $X^p = \langle X_1^p, X_2^p, \dots, X_n^p \rangle$ of its coordinates, we establish a 1-1 correspondence between the points of A and the points of $A_{\mathfrak{F}}$. (See the definitions of $A_{\mathfrak{F}}$ and \oplus , Theorem 2.8 and Lemma 2.9 (ii).) It remains to be shown that this correspondence establishes an isomorphism between \mathfrak{M} and $\mathfrak{K}_n(\mathfrak{F})$. This reduces to showing that the relations B and D among points of A can be characterized in terms of the coordinates of these points in exactly the same way in which the relations $B_{\mathfrak{F}}$ and $D_{\mathfrak{F}}$ among points of $A_{\mathfrak{F}}$ have been defined in the Klein model $\mathfrak{K}_n(\mathfrak{F})$.

Consider two distinct points p and q in A and the correlated n -tuples of coordinates X^p and X^q . We first express the free segment $[pq]$ in terms of X^p and X^q . An easy but lengthy calculation, based exclusively upon Lemmas 2.9(i), 2.10, and 2.13, leads to

$$(5) \quad (C[pq])^2 = \Psi(X^p, X^q),$$

where Ψ is the function used in describing the Klein space. (The argument is analogous to that used in Euclidean case, with the difference that rectangles are replaced by right quadrangles.) From (5), Lemma 2.9(iii), and Lemma 2.14 (i) we conclude at once that the condition

$$\Psi(X^p, X^r) = \frac{\Psi(X^p, X^q) \cdot \Psi(X^q, X^r)}{(1 + \sqrt{1 - \Psi(X^p, X^q)} \cdot \sqrt{1 - \Psi(X^q, X^r)})^2}$$

is necessary and sufficient for points p, q, r to satisfy the formula $B(pqr)$. Similarly, from (5) and Lemma 2.14 (ii) we conclude that the condition

$$\Psi(X^p, X^q) = \Psi(X^r, X^s)$$

is necessary and sufficient for points p, q, r, s to satisfy the formula $D(pqrs)$. Thus the proof is completed.

Using Theorem 2.15 instead of 1.1 we obtain of course a new proof of Theorem 1.2 and, actually, we can extend this result to arbitrary dimension n :

THEOREM 2.16. *The theory \mathcal{H}_n ($n \geq 2$) is complete and decidable but not finitely axiomatizable.*

3. Applications to Related Geometrical Systems. Using the main results stated in [5] for Euclidean geometry and in this paper for hyperbolic geometry we shall now establish fundamental metamathematical properties of elementary absolute geometry. The discussion in [5] has been restricted to the two-dimensional case only for simplicity of formulation and the results established there clearly extend to elementary n -dimensional Euclidean geometry \mathcal{E}_n for any $n \geq 2$.

Let the formalized system \mathcal{A}_n of n -dimensional absolute geometry be a theory which has the same symbolism as \mathcal{E}_n and \mathcal{H}_n and the axiom system of which is obtained by omitting Euclid's axiom, A8, in the axiom system of \mathcal{E}_n (or the negation of A8 in the axiom system of \mathcal{H}_n). Thus a sentence is valid in \mathcal{A}_n if and only if it is valid in both \mathcal{E}_n and \mathcal{H}_n . As simple consequence of Theorem 1 in [5] and Theorem 2.15 in the present paper we obtain.

THEOREM 3.1. *\mathfrak{M} is a model of \mathcal{A}_n ($n \geq 2$) if and only if it is isomorphic either with the Cartesian space $\mathbb{C}_n(\mathfrak{F})$ or with the Klein space $\mathbb{K}_n(\mathfrak{F})$ over some real closed field \mathfrak{F} .*

Theorem 3.1 contains a description of all models of \mathcal{A}_n which is however not uniform in its character; the class of models proves to consist of two widely different subclasses. It would be interesting to obtain a more homogeneous characterization of this class.

Theorems 2, 3, and 4 in [5] and Theorem 2.16 in the present paper imply the following theorems 3.2–3.4 as direct corollaries.

THEOREM 3.2. *The theory \mathcal{A}_n ($n \geq 2$) has just two complete and consistent extensions, in fact, \mathcal{E}_n and \mathcal{H}_n .*

A consequence of Theorem 3.2. is that Euclid's axiom can be equivalently replaced in the axiom system of \mathcal{E}_n by any sentence whatsoever which is valid in \mathcal{E}_n but not in \mathcal{H}_n ; the same of course applies to the negation of the Euclid's axiom in the axiom system of \mathcal{H}_n .

THEOREM 3.3. *The theory \mathcal{A}_n ($n \geq 2$) is decidable.*

This theorem is an improvement of Tarski's decision theorem for \mathcal{E}_n .

THEOREM 3.4. *The theory \mathcal{A}_n ($n \geq 2$) is not finitely axiomatizable.*

In conclusion we wish to make some remarks concerning the system $\overline{\mathcal{H}}_n$ of *non-elementary n -dimensional hyperbolic geometry*. The main difference between the symbolisms of \mathcal{H}_n and $\overline{\mathcal{H}}_n$ consists primarily in the fact that all the variables occurring in the former range over points, while the latter contains also variables ranging over arbitrary point sets. (The question whether $\overline{\mathcal{H}}_n$ contains in addition variables of higher orders ranging over families of sets, etc. is irrelevant for the subsequent remarks.) The axiom system of $\overline{\mathcal{H}}_n$ is obtained from that of \mathcal{H}_n by replacing the infinite collection of elementary continuity axioms by one non-elementary axiom (see [5], p. 18). In every model \mathfrak{M} of $\overline{\mathcal{H}}_n$ the ordered field \mathfrak{F} in which the system \mathfrak{S} can be imbedded (see Theorem 2.8) proves to be continuously ordered. Since a continuously ordered field \mathfrak{F} is isomorphic with the field \mathfrak{R} of real numbers, the correlated Klein space $\mathfrak{K}_n(\mathfrak{F})$ is isomorphic with the Klein space $\mathfrak{K}_n(\mathfrak{R})$. Thus, by Theorem 2.15, we conclude that every model \mathfrak{M} of $\overline{\mathcal{H}}_n$ is isomorphic with the Klein model $\mathfrak{K}_n(\mathfrak{R})$. In this way we arrive at

THEOREM 2.18. *The theory $\overline{\mathcal{H}}_n$ is categorical.*

This result is well known but all other proofs which are known to the author are based upon an analytic formula for $\Pi(X)$ (see page 34) and hence upon some properties of exponential and trigonometric functions.⁶

⁶ While the paper was in press the author noticed that a direct two-dimensional proof of Lemma 2.4 (cf. Note 2.7 on p. 44) results at once from a theorem due to Liebmann in [6], p. 191.

Moreover, the author succeeded in constructing in \mathcal{A}_n ($n \geq 2$) an *absolute calculus of segments*. This calculus leads to the representation theorems for both Euclidean and Bolyai-Lobachevskian geometries.

Bibliography

- [1] HILBERT, D. *Grundlagen der Geometrie*. 8th ed., Stuttgart 1956.
- [2] HJELMSLEV, J. *Beiträge zur Nicht-Eudoxischen Geometrie* I–II. Det. Kgl. Danske Videnskabernes Selskab, Matematisk-Fysiske Meddelelser, vol. 21 (1944), Nr. 5.
- [3] SZÁSZ, P. *Direct introduction of Weierstrass homogeneous coordinates in the hyperbolic plane, on the basis of the endcalculus of Hilbert*. This volume, pp. 97–113.
- [4] TARSKI, A. *A decision method for elementary algebra and geometry*. 2nd ed., Berkeley and Los Angeles 1951.
- [5] ——— *What is elementary geometry?* This volume, pp. 16–29.
- [6] LIEBMANN, H., *Elementargeometrischer Beweis der Parallelenkonstruktion und neue Begründung der trigonometrischen Formeln der hyperbolischen Geometrie*, Mathematische Annalen, vol. 61 (1905), pp. 185–199.

DIMENSION IN ELEMENTARY EUCLIDEAN GEOMETRY ¹

DANA SCOTT

Princeton University, Princeton, New Jersey, U.S.A.

Introduction. It has been well over one hundred years since higher dimensional geometry made its appearance in mathematics and at least fifty years since the terminology of infinite dimensional spaces came into general use. No one can deny the enrichment of the subject brought about by the introduction of these notions, but even though the infinite dimensional spaces would seem to be a direct generalization of the finite dimensional spaces, it is clear that their importance in mathematics really lies in a different direction. In finite dimensions we are concerned with ever more complicated configurations of points, lines, planes, spheres, or other algebraic varieties, and in this study a knowledge of facts in higher dimensions often leads to a better understanding of the lower dimensions. Of course, all such configurations are possible in an infinite dimensional space, but in the study of any one particular problem so little of the space is used that one might as well work in only a finite number of dimensions. Thus, the question arises whether there is really anything new in infinite dimensional geometry. The applications of infinite dimensional geometry to analysis and the study of function spaces are something new and beyond the finite dimensional theory, but this is not what is meant. From the standpoint of pure geometry, is there anything new? In particular, are there different kinds of infinite dimensional spaces? Of course, anyone can think of two distinct Hilbert spaces, for example, but is there any geometrical property that distinguishes them? Cardinality is a property that will often distinguish between two infinite dimensional spaces; however, the cardinal number of a set is not really associated with the internal structure of the space in isolation but only becomes meaningful in comparisons with other sets. Thus, in the study of geometrical properties of spaces we wish to restrict attention to those constructions that can be carried out within the space itself making use of only the given geometrical notions. This point of view seems even to throw doubt on

¹ The results of this paper represent a portion of a thesis submitted to the Faculty of Princeton University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

topological questions. The most useful facts of point-set topology nearly always rest on operations performed on arbitrary subsets of the space, and since the time of Cantor we have realized how vastly complicated these subsets may become. Indeed, point-set topology with its heavy use of infinite combinations and infinite repetitions of operations, though derived from geometrical intuition, is a totally new discipline that has moved far from the special world of Euclidean geometry. The same may be said for other questions of the analysis of Hilbert spaces such as completeness, existence of orthonormal bases, and the like. Thus, we may be led to the conclusion that the usual geometrical notions do not involve infinite sets or infinite sequences of points, and this will be the convention adapted in the present paper. If the reader does not entirely agree with this point of view, at least it is hoped that he will agree that *elementary* geometrical notions do not involve infinite sets and that he admits that even the infinite dimensional spaces contain the material for many elementary constructions, so that there is a meaningful question whether infinite dimensional spaces can be distinguished by their elementary properties.

First of all it must be said what Euclidean spaces, finite or infinite dimensional, actually are. The definition chosen in Section 1 is the standard one making use of vector spaces. Geometrical properties of spaces must be formulated in terms of geometrically meaningful notions. In the case of elementary properties there is no loss of generality in considering only finitary relations between points, and in this context the term *geometrically meaningful relation* or simply *geometrical relation* is given a precise definition. Finally elementary geometrical properties are identified with those properties of a space expressible in sentences of the first-order predicate logic in terms of the geometrical relations over the space. Before giving the specifically geometric results, a general theorem in the theory of models of the first-order logic is presented in Section 2. The general result is then applied in a straight forward way to geometry in Section 3, and it is shown that *there are no elementary geometrical properties distinguishing any two infinite dimensional Euclidean spaces*. In particular, for any given formal property, a very simple method is given for calculating a finite dimension, m say, such that the property is true in spaces of dimension m if and only if it is true in all higher dimensions including all the infinite dimensions. The consequences of this state of affairs for a certain formal theory of geometry are indicated in Section 4.

The author would like to thank Professor Tarski, who originally proposed the problem and who made many helpful comments on the formulation of the results.

1. Euclidean Spaces and Geometrical Relations. Before discussing any formal theory, it is necessary to determine the standard domains of discourse to which the theory will be applied. As regards geometry, if we were concerned only with finite dimensions, we could think simply of the ordinary n -dimensional cartesian spaces whose points are n -tuples of real numbers. But these are not sufficient for our purposes. In any case, there is no need to think of a particular coordinate system as in the cartesian spaces, because a distinguished coordinate system is not a purely geometrical notion. A definition by vector space methods solves the problem and eliminates any distinguished set of coordinates. In the first place, a (*standard*) *Euclidean space* will be a vector space over the field of real numbers having any finite or infinite linear dimension. In addition, to give the essential Euclidean character to the space, a notion sufficient for questions of distance and perpendicularity has to be supplied. A positive definite *inner product* on the space will do just that. To sum up, a *Euclidean space* is a 4-tuple $\langle V, +, \cdot, \cdot \rangle$, where V is a set of elements called the points of the space; $\langle V, + \rangle$ is an abelian group; \cdot is an operation from the cartesian product of the real numbers with V to the set V satisfying the following properties for all reals α, β and all $x, y \in V$:

- (i) $1 \cdot x = x$;
- (ii) $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x$;
- (iii) $(\alpha + \beta) \cdot x = (\alpha \cdot x) + (\beta \cdot x)$;
- (iv) $\alpha \cdot (x + y) = (\alpha \cdot x) + (\alpha \cdot y)$;

and finally \cdot is an operation from pairs of elements of V to real numbers such that for all reals α, β and all $x, y, z \in V$:

- (v) If $x \neq 0$, then $x \cdot x > 0$;
- (vi) $x \cdot y = y \cdot x$;
- (vii) $((\alpha \cdot x) + (\beta \cdot y)) \cdot z = \alpha(x \cdot z) + \beta(y \cdot z)$;

where the symbol 0 denotes in the hypothesis of (v) the zero element of the group.

German capital letters \mathfrak{B} and \mathfrak{B} will be used to denote Euclidean

spaces, and the corresponding set of points will be denoted by Roman capitals V and W .

In a Euclidean space the *distance* between points x and y , in symbols $\|x - y\|$, can be introduced by definition:

$$\|x - y\| = + \sqrt{(x - y) \cdot (x - y)},$$

where $(x - y)$ on the right hand side means $(x + ((-1) \cdot y))$.

The above treatment of Euclidean spaces, though it does not involve a choice of a particular coordinate system, does involve using a distinguished point: the origin or zero vector 0 . The dependence on 0 is eliminated in our definitions of the notions of subspace and isometry. A *subspace* of a Euclidean space \mathfrak{B} is a non-empty subset X of V such that whenever x and y are points in X , then $\alpha \cdot x + (1 - \alpha) \cdot y$ is in X for all real numbers α . In other words, if a subspace contains two points of a line, then it must contain all points of the line. An *isometry* from a subspace X of a Euclidean space \mathfrak{B} onto a subspace Y of a Euclidean space \mathfrak{B} is a one-one function f from X onto Y preserving the distance between points; that is, if x and y are in X , then $\|x - y\| = \|f(x) - f(y)\|$, where the distance on the left hand side of the formula refers to the space \mathfrak{B} , and, on the right hand side, to the space \mathfrak{B} .

In this terminology, subspaces of a Euclidean space are not again Euclidean spaces since they are not necessarily vector subspaces. If a subspace contains the zero vector, then it is a vector subspace. It is thus obvious that a translation of the space will always carry any subspace onto a vector subspace, and hence we can say that every subspace of a Euclidean space is isometric with a Euclidean space. Clearly, isometries between Euclidean spaces always preserve dimension, and so every subspace of a Euclidean space has an unambiguous dimension.

Having thus defined the dimension of a subspace, a simple property of subspaces that is needed in the later work can be stated:

LEMMA 1.1. *Every set of $m + 1$ points of a Euclidean space is contained in a subspace of dimension at most m .*

Somewhat more complicated but very easy to prove is the following:

LEMMA 1.2. *Let X and Y be two subspaces of a Euclidean space \mathfrak{B} having the same finite dimension. Then there is an isometry of \mathfrak{B} onto itself, mapping X onto Y , and leaving the intersection $X \cap Y$ pointwise fixed.*

The question we turn to next is that of defining the concept of a geometrically meaningful relation between points. There are two different aspects to the question. First, we can consider a fixed Euclidean space and relations between the points of that one space. Second, the class of all Euclidean spaces can be considered, and relations in different spaces can be compared. For the first problem the answer is very simple: In a Euclidean space \mathfrak{B} , the *geometrical relations* over \mathfrak{B} are just those relations between points of V invariant under the group of all isometries of \mathfrak{B} onto itself. In other words, an n -ary relation R over V , or a subset R of the cartesian power V^n , is a *geometrical relation* if for all isometries f of \mathfrak{B} onto itself and all n -tuples $\langle x_0, \dots, x_{n-1} \rangle \in V^n$ we have $\langle x_0, \dots, x_{n-1} \rangle \in R$ if and only if $\langle f(x_0), \dots, f(x_{n-1}) \rangle \in R$. The above definition, of course, agrees with the well-known program of Klein which asserts that the group of motions of the space should determine the geometry.

When we pass over to the class of all Euclidean spaces some care must be taken. There are indeed some set-theoretical problems connected with the idea of the class of all spaces. These problems do not cause any essential difficulties and can all be solved by adopting a standard type of formal set-theoretical framework. Rather more important here is the question of comparison of different spaces. A geometrical relation over the class of all Euclidean spaces should be an assignment of one geometrical relation to each particular space. Clearly isometric spaces should get isometric relations; but more than this, the assignment should be insensitive to dimension. It would be hopeless to try to classify all ways of assigning one kind of relation to one-dimensional spaces, another kind to two-dimensional spaces, a third to three dimensions and so on. So we are lead to the restriction that a geometrical relation over all spaces is to be invariant under isometries not only from one space *onto* another, but also from one space *into* another. In more formal terms: an n -ary *geometrical relation* over the class of all Euclidean spaces is a function R that assigns to each space \mathfrak{B} a subset $R_{\mathfrak{B}}$ of V^n such that if f is an isometry of \mathfrak{B} into a space \mathfrak{B} , then for all n -tuples $\langle x_0, \dots, x_{n-1} \rangle \in V^n$ we have $\langle x_0, \dots, x_{n-1} \rangle \in R_{\mathfrak{B}}$ if and only if $\langle f(x_0), \dots, f(x_{n-1}) \rangle \in R_{\mathfrak{B}}$. The effect of the above definition is to assure that if X is a subspace of a Euclidean space \mathfrak{B} , then the relation $R_{\mathfrak{B}} \cap X^n$, the restriction to X , is isometric to the relation obtained by considering X a Euclidean space in itself.

There are many examples of geometrical relations. The first interesting case is that of binary relations. It can be shown that there are $2^{2^{\aleph_0}}$ geometrical binary relations; in fact, they can all be obtained in the following

way: Let D be any set of non-negative real numbers. For each space \mathfrak{B} , let the relation $R_{\mathfrak{B}}$ be defined by the condition $\langle x, y \rangle \in R_{\mathfrak{B}}$ if and only if $\|x - y\| \in D$, for all $x, y \in V$. Then R is a geometrical relation. For ternary relations we need mention only a few: betweenness, being the midpoint, collinearity, forming an equilateral triangle, being equidistant from two points, and so on. A similar description of all such relations can easily be given in terms of sets of triples of real numbers.

Finally it is to be noted that the definition can be extended to cover geometrical relations between lines, planes, spheres, and the like, but this is hardly ever necessary and will not be considered here. In view of the fact that the various algebraic loci are completely determined by a finite number of points lying on them, any relation between such objects can be encoded into an equally powerful relation between points. For example, a binary relation between lines can always be replaced by a quaternary relation between points. Also there is no need to consider more than one geometrical relation between points, since two relations, one n -ary and one m -ary say, can always be replaced by a single $(n + m)$ -ary relation in an obvious way.

2. Arithmetical Extensions of Finite Degree. All terminology of the paper of Tarski and Vaught [2] will be adopted for the purposes of this section, except that relational systems $\langle A, R \rangle$ are considered where R is n -ary relation, or relation of rank n , rather than just a ternary relation. The integer n , however, is to be fixed for the discussion. The formal theory T in the first-order predicate logic must then contain an n -placed predicate symbol P , as well as the standard logical symbols.

In particular, we are interested in the specific algebraic condition given by Tarski and Vaught [2] in Theorem 3.1 for one relational system to be an arithmetical extension of another. The general notion of arithmetical extension defined in that paper concerns all possible formulas of the first-order logic, and for the purposes of this paper a weaker notion involving only a restricted class of formulas is needed. The formal definition follows.

DEFINITION 2.1. *The system $\mathfrak{S} = \langle B, S \rangle$ is called an m -degree arithmetical extension of the system $\mathfrak{R} = \langle A, R \rangle$ if the following two conditions are satisfied:*

- (i) \mathfrak{S} is an extension of \mathfrak{R} ;

- (ii) *for every formula ϕ containing at most m distinct variables and every sequence $x \in A^{(\omega)}$, x satisfies ϕ in \mathfrak{R} if and only if x satisfies ϕ in \mathfrak{S} .*

It should be noted that there is no loss of generality in condition (ii) if the formula ϕ is required to contain only variables from the specific list v_0, v_1, \dots, v_{m-1} , and then the sequence x can be chosen simply from the set A^m .

A generalization of Theorems 3.1 of [2] can now be stated and proved.

THEOREM 2.2. *The following two conditions are (jointly) sufficient for a system $\mathfrak{S} = \langle B, S \rangle$ to be an m -degree arithmetical extension of a system $\mathfrak{R} = \langle A, R \rangle$:*

- (i) *\mathfrak{S} is an extension of \mathfrak{R} ;*
- (ii) *for any subset A' of A with less than m elements and any element b of B , there exists an automorphism f of \mathfrak{S} such that f leaves A' pointwise fixed and $f(b) \in A$.*

PROOF. Using the remark following Definition 2.1 we restrict attention to formulas involving at most the variables v_0, v_1, \dots, v_{m-1} and proceed by induction on the length of formulas. Assuming conditions (i) and (ii) above, the following is the statement to be proved for formulas ϕ of the restricted type:

- (*) *for all $x \in A^m$, x satisfies ϕ in \mathfrak{R} if and only if x satisfies ϕ in \mathfrak{S} .*

The statement (*) is obviously true for atomic formulas, and it is very easy to show that if (*) holds for formulas ϕ and ψ then it holds for $\neg \phi$ and $\phi \wedge \psi$. Suppose now that (*) holds for ϕ , and consider the formula $\bigvee v_k \phi$, where $k < m$. Assume first that $x \in A^m$ and x satisfies $\bigvee v_k \phi$ in \mathfrak{R} . Then for some element $a \in A$, $x(k/a)$ satisfies ϕ in \mathfrak{R} . By the hypothesis $x(k/a)$ satisfies ϕ in \mathfrak{S} , and hence x satisfies $\bigvee v_k \phi$ in \mathfrak{S} , as was to be shown. Assume now that $x \in A^m$ and x satisfies $\bigvee v_k \phi$ in \mathfrak{S} . Let $b \in B$ such that $x(k/b)$ satisfies ϕ in \mathfrak{S} . The set $A' = \{x_i | i < m, i \neq k\}$ is a subset of A with fewer than m elements. In view of condition (ii), let f be an automorphism of \mathfrak{S} leaving A' pointwise fixed and with $f(b) \in A$. Since f is an automorphism of \mathfrak{S} , the sequence $\langle f(x_0), \dots, f(x_{k-1}), f(b), \dots, f(x_{m-1}) \rangle = x(k/f(b))$ must satisfy ϕ in \mathfrak{S} . Hence, from (*) for ϕ , we have that $x(k/f(b))$ satisfies ϕ in \mathfrak{R} , and finally x satisfies $\bigvee v_k \phi$ in \mathfrak{R} , which completes the proof that (*) holds for $\bigvee v_k \phi$. Thus, by induction (*) is established for all formulas and the theorem is proved.

In the original version of this paper, the author proved a somewhat different form of Theorem 2.2 which does not require the existence of automorphisms of the system \mathfrak{S} but rather uses a whole class of isomorphic subsystems \mathfrak{B} of \mathfrak{S} whose union covers the set B . However, Euclidean spaces possesses so many isometries, as was noted in Lemma 1.2 of Section 1, that the simpler theorem just given is quite adequate for the results of the next section. The other version of the general algebraic condition for m -degree extensions and its applications will be published elsewhere. Notice that Theorem 2.2 implies Theorem 3.1 of Tarski-Vaught [2], since being an arithmetical extension is equivalent to being an m -degree extension for each m , and the condition (ii) of their Theorem 3.1 obviously implies conditions (ii) of Theorem 2.1 above.

3. Relational Systems Derived from Euclidean Spaces. Let \mathfrak{B} be a Euclidean space and let R be an n -ary geometrical relation over \mathfrak{B} . The system $\langle V, R \rangle$ is a relational system and any subspace X of \mathfrak{B} yields a corresponding subsystem $\langle X, R \cap X^n \rangle$. Our first theorem shows the relation of the theory of first-order sentences true of R in the whole space \mathfrak{B} to those true in the subspace X .

THEOREM 3.1. *If R is an n -ary geometrical relation over the Euclidean space \mathfrak{B} and X is a subspace of \mathfrak{B} of dimension at least m , then the relational system $\langle V, R \rangle$ is an $(m+1)$ -degree arithmetical extension of $\langle X, R \cap X^n \rangle$.*

PROOF. We need only verify condition (ii) of Theorem 2.2. Let X' be a subset of X with at most m elements. Since we may obviously assume $m > 0$, X' can be contained in a subspace Y of dimension $m-1$ which is also contained in X . Let Y_0 be a subspace of dimension exactly m containing Y and contained in X . Let b be any point in V not in X . Obviously we can find a subspace Y_1 of dimension exactly m containing Y and containing b . Since Y_0 is included in X and b is not, $Y_0 \cap Y_1 = Y$. Using Lemma 1.2, let f be an isometry of \mathfrak{B} onto itself taking Y_1 onto Y_0 and leaving Y pointwise fixed. The function f will thus be an automorphism of $\langle V, R \rangle$ such that $f(b) \in X$, which completes the proof.

COROLLARY 3.2. *If R is an n -ary geometrical relation over the Euclidean space \mathfrak{B} and X is an infinite dimensional subspace of \mathfrak{B} , then the relational system $\langle V, R \rangle$ is an arithmetical extension of $\langle X, R \cap X^n \rangle$.*

COROLLARY 3.3. *If R is an n -ary geometrical relation over the Euclidean space \mathfrak{B} and X and Y are two infinite dimensional subspaces of \mathfrak{B} , then the relational systems $\langle X, R \cap X^n \rangle$ and $\langle Y, R \cap Y^n \rangle$ are arithmetically equivalent.*

In less formal terms the above results can be explained in the following way. Let \mathfrak{B} be an Euclidean space and R be a geometrical relation. Consider a sentence ϕ in the formal first-order theory of a predicate that is to be interpreted as the relation R . We ask whether ϕ expresses a true property of R . Now ϕ contains only finitely many symbols and in particular only a finite number of variables, $m + 1$ say. Theorem 3.1 shows us that the truth of ϕ can be established by looking not at the whole space, but only at m -dimensional subspaces of \mathfrak{B} . If \mathfrak{B} were already of a dimension smaller than m , this result is not of much help. However, if \mathfrak{B} has a very large dimension or is infinite dimensional, then the reduction is considerable. In particular, Corollary 3.3 shows that no single first-order property or even a set of first-order properties of the geometrical relation R can ever distinguish between two infinite dimensional subspaces of \mathfrak{B} .

We turn now from one space to the class of all spaces. Here we need to consider geometrical relations over the whole class of spaces as defined in Section 1. As a direct consequence of Theorem 3.1 we obtain:

THEOREM 3.4. *If R is a geometrical relation over the class of all Euclidean spaces and ϕ is a sentence of first-order logic with $(m + 1)$ distinct variables, then ϕ is true in all relational systems $\langle V, R_{\mathfrak{B}} \rangle$, where \mathfrak{B} is a Euclidean space of dimension at least m , if and only if ϕ is true in at least one such relational system.*

COROLLARY 3.5. *If R is a geometrical relation over the class of all Euclidean spaces and \mathfrak{B} and \mathfrak{B}' are infinite dimensional Euclidean spaces, then the relational systems $\langle V, R_{\mathfrak{B}} \rangle$ and $\langle W, R_{\mathfrak{B}'} \rangle$ are arithmetically equivalent.*

The argument that leads to 3.5 can be extended to show that there is no collection of geometrical relations and no collection of their first-order properties that can distinguish between any two infinite dimensional Euclidean spaces. It should be clear from the proof given above that if we only wanted this result about infinite dimensional spaces, it would be possible to use Theorem 3.1 of Tarski-Vaught [2] directly without going through the generalization of that method given in Section 2. However,

the relation between truth in the whole space and truth in its finite dimensional subspaces as developed here in Theorem 3.4 leads to an even stronger result about infinite dimensional geometry as is explained in the next section. Furthermore, it allows us to establish a criterion for determining whether a relation defined in first-order logic in terms of a given geometrical relation is again such a relation. This criterion is presented in Theorem 3.6 below.

First it must be made clear when one relation is definable (in first-order logic) in terms of another relation. Let R be a geometrical relation and let φ be a formula in the first order theory of R whose free variables are all contained in the list v_0, v_1, \dots, v_{p-1} . We can easily think of φ as defining a new p -ary relation, S say, in terms of R . Of course this definition must be made relative to each Euclidean space separately, and so S must be thought of as a function from spaces \mathfrak{B} to subsets $S_{\mathfrak{B}}$ of V^p . Finally in precise terms we say that S is *defined* by φ in terms of R if for each Euclidean space \mathfrak{B} and for all sequences $x \in V^{(\omega)}$, we have $\langle x_0, \dots, x_{p-1} \rangle \in S_{\mathfrak{B}}$ if and only if x satisfies φ in $\langle V, R_{\mathfrak{B}} \rangle$. Not every formula φ leads to a geometrical relation S , however. To see this, let φ contain freely none of the variables v_0, \dots, v_{p-1} , and choose the formula in such a way that it expresses a property of spaces true in only one dimension, then S will not be geometrical. The test that S must pass to be a geometrical relation is given next.

THEOREM 3.6. *Let the p -ary relation S be defined by the formula φ in terms of the geometrical relation R . Suppose further that the total number of variables in φ , including the free variables, is $m + 1$. Then S is a geometrical relation if and only if for all Euclidean spaces \mathfrak{B} and \mathfrak{B} of dimension at most m and all isometries f of \mathfrak{B} into \mathfrak{B} , we have for all sequences $x \in V^{(\omega)}$, $\langle x_0, \dots, x_{p-1} \rangle \in S_{\mathfrak{B}}$ if and only if $\langle f(x_0), \dots, f(x_{p-1}) \rangle \in S_{\mathfrak{B}}$.*

PROOF. Obviously, if S is a geometrical relation, then it satisfies the condition given for isometries. Suppose then that S is not a geometrical relation. Thus, there must be Euclidean spaces \mathfrak{B} and \mathfrak{B} and an isometry f from \mathfrak{B} into \mathfrak{B} and a sequence $x \in V^{(\omega)}$ such that the formulas $\langle x_0, \dots, x_{p-1} \rangle \in S_{\mathfrak{B}}$ and $\langle f(x_0), \dots, f(x_{p-1}) \rangle \in S_{\mathfrak{B}}$ are not equivalent. By the symmetry of the situation we need treat only the case where $\langle x_0, \dots, x_{p-1} \rangle \in S_{\mathfrak{B}}$ and $\langle f(x_0), \dots, f(x_{p-1}) \rangle \notin S_{\mathfrak{B}}$. Since ϕ defines S , we conclude that x satisfies ϕ in $\langle V, R_{\mathfrak{B}} \rangle$ and the sequence $\langle f(x_0), f(x_1), \dots \rangle$ does not satisfy ϕ in $\langle W, R_{\mathfrak{B}} \rangle$. Due to the fact that ϕ has only free variables in the set $\{v_0, \dots, v_{p-1}\}$ we can assume without loss of generality

that $\{x_i | i < \omega\} = \{x_0, \dots, x_{p-1}\}$. Now by hypothesis $p \leq m + 1$, and so there exists a subspace X of \mathfrak{B} of dimension at most m containing the set $\{x_0, \dots, x_{p-1}\}$; in particular, if \mathfrak{B} is of dimension at most m , we shall assume $X = V$, and otherwise that X is of dimension exactly m . In any case we can conclude with the aid of Theorem 3.1 that the sequence x satisfies ϕ in $\langle X, R_{\mathfrak{B}} \cap X^p \rangle$. The image of X under f is a subspace X' of \mathfrak{B} of dimension equal to that of X . Let Y be a subspace of \mathfrak{B} that is either equal to W in case \mathfrak{B} is of dimension less than m or is of dimension exactly m , and which in any case contains X' . By the same argument as above the sequence $\langle f(x_0), f(x_1), \dots \rangle$ does not satisfy ϕ in $\langle Y, R_{\mathfrak{B}} \cap Y^p \rangle$. Now the two subspaces X of \mathfrak{B} and Y of \mathfrak{B} are themselves isometric with Euclidean spaces \mathfrak{B}' and \mathfrak{B}'' of dimensions at most m by isometries g and h where g is from V' onto X and h is from Y onto W' . Let $f' = hfg$, which will be an isometry from \mathfrak{B}' into \mathfrak{B}'' . By our very construction we can obviously conclude that the sequence $\langle g^{-1}(x_0), g^{-1}(x_1), \dots \rangle$ satisfies ϕ in $\langle V', R_{\mathfrak{B}'} \rangle$ and hence $\langle g^{-1}(x_0), \dots, g^{-1}(x_{p-1}) \rangle \in S_{\mathfrak{B}'}$, while $\langle hf(x_0), \dots, hf(x_{p-1}) \rangle \notin S_{\mathfrak{B}''}$. This finally shows that S does not satisfy the condition of the theorem, which completes the proof.

4. Axiomatic Geometry. In his paper [3], Tarski presents a particularly neat axiomatic system for two-dimensional Euclidean geometry in terms of the basic notions of *betweenness* and *equidistance*. As is indicated in [3] an axiomatization for any finite dimension can be obtained by a simple change in two of the axioms. All these axiomatic theories are decidable, and it follows from the method in Tarski's monograph [1] that there is even a uniform method for deciding for each integer m whether an elementary sentence in terms of betweenness and equidistance is true in Euclidean spaces of dimension m . It is to be shown here that there is also an effective decision method for the class of sentences true in infinite dimensional spaces.

Let B and E be respectively ternary and quaternary geometrical relations denoting the *betweenness* and *equidistance* relations in Euclidean spaces. The first-order theory, then, must contain a ternary and a quaternary predicate symbol. Let \mathcal{E}_m , $m < \omega$, be the class of all sentences of this first-order theory true in the relational systems $\langle V, B_{\mathfrak{B}}, E_{\mathfrak{B}} \rangle$ where \mathfrak{B} is an m -dimensional Euclidean space. Since all m -dimensional spaces are isometric, the theory \mathcal{E}_m is complete. In this section we shall often use the word *theory* to mean any class of sentences of the first-order logic that is consistent and is closed under all the usual rules of deduction; while a

complete theory is a maximal such class. Let $\mathcal{E} = \bigcap_{m < \omega} \mathcal{E}_m$ be the common part of all these theories, that is, the class of sentences true in all finite dimensions. \mathcal{E} is, of course, not a complete theory, but it is a decidable theory as will be shown below. One further theory will be considered, namely $\mathcal{E}_\infty = \bigcup_{m < \omega} \bigcap_{n > m} \mathcal{E}_n$, that is, the class of sentences true in all but a finite number of dimensions. \mathcal{E}_∞ is a theory since it is the union of an increasing sequence of theories, but what is surprising is that \mathcal{E}_∞ is a complete theory and, in fact, is the class of sentences true in all infinite dimensions. We turn now to the systematic account of these results.

LEMMA 4.1. $\mathcal{E}_m \sim \bigcup_{n \neq m} \mathcal{E}_n \neq 0$

PROOF. In words: there is a sentence true in the dimension m but not true in any other dimension. To demonstrate this one has only to translate into formal logical symbols the sentence that expresses the fact that there exists a configuration of $m + 1$ distinct and mutually equidistant points, but no such configuration with $m + 2$ points. Notice that the trivial dimension $m = 0$ is accomodated quite nicely.

LEMMA 4.2. If Δ_m is any sentence in the set $\mathcal{E}_m \sim \bigcup_{n \neq m} \mathcal{E}_n$, then $\mathcal{E}_m = cl(\mathcal{E} \cup \{\Delta_m\})$.

REMARK. The symbol $cl(\mathcal{X})$ denotes the closure of the set of sentences \mathcal{X} under the rules of deduction of the first-order predicate logic. Thus 4.2 expresses the fact that the theory \mathcal{E}_m results from the theory \mathcal{E} by the addition of any single axiom chosen as indicated in the hypothesis of the lemma.

PROOF. Assuming the hypothesis, let ϕ be any sentence in \mathcal{E}_m and consider the implication $[\Delta_m \rightarrow \phi] = \neg[\Delta_m \wedge \neg \phi]$. Clearly $[\Delta_m \rightarrow \phi] \in \mathcal{E}_m$ since $\phi \in \mathcal{E}_m$. But also, for any $n \neq m$, $\Delta_m \notin \mathcal{E}_n$ and so $\neg \Delta_m \in \mathcal{E}_n$; hence, $[\Delta_m \rightarrow \phi] \in \mathcal{E}_n$. It follows at once that $[\Delta_m \rightarrow \phi] \in \mathcal{E}$. This argument shows that $\mathcal{E}_m \subseteq cl(\mathcal{E} \cup \{\Delta_m\})$. The obviousness of the opposite inclusion completes the proof.

THEOREM 4.3. The only finite complete extensions of the theory \mathcal{E} are the theories \mathcal{E}_m , $m < \omega$.

PROOF. That each complete theory \mathcal{E}_m is a finite extension of \mathcal{E} is the content of Lemmas 4.1 and 4.2. Assume then that \mathcal{E}_* is a finite complete extension of \mathcal{E} with $\mathcal{E}_* \neq \mathcal{E}_m$ for all $m < \omega$. Let Δ_* be the single axiom

needed to have $\mathcal{E}_* = cl(\mathcal{E} \cup \{\Delta_*\})$. Since $\mathcal{E}_* \neq \mathcal{E}_m$, it follows that $\Delta_* \notin \mathcal{E}_m$. Thus, $\neg \Delta_* \in \mathcal{E}_m$, for all $m < \omega$, which implies $\neg \Delta_* \in \mathcal{E}$. Thus, the theory \mathcal{E}_* would have to be inconsistent, which is impossible.

LEMMA 4.4. *If the sentences Δ_m are chosen as in Lemma 4.2, then $\mathcal{E}_\infty = cl(\mathcal{E} \cup \{\neg \Delta_m | m < \omega\})$.*

PROOF. $\neg \Delta_m \in \bigcap_{n > m} \mathcal{E}_n$ by construction, and hence $\neg \Delta_m \in \mathcal{E}_\infty$ for each $m < \omega$. Thus, $cl(\mathcal{E} \cup \{\neg \Delta_m | m < \omega\}) \subseteq \mathcal{E}_\infty$. Let ϕ be any sentence in \mathcal{E}_∞ . There exists an integer m such that $\phi \in \bigcap_{n > m} \mathcal{E}_n$. Consider the implication $[(\neg \Delta_0 \wedge \neg \Delta_1 \wedge \dots \wedge \neg \Delta_{m-1}) \rightarrow \phi]$. It is easy to see that this sentence is in \mathcal{E} and hence $\phi \in cl(\mathcal{E} \cup \{\Delta_m | m < \omega\})$. The converse inclusion is thus established.

LEMMA 4.5. *\mathcal{E}_∞ is the set of sentences true in all infinite dimensional spaces and is complete.*

PROOF. This lemma is a direct consequence of Theorem 3.4, Corollary 3.5, and the definition of \mathcal{E}_∞ .

THEOREM 4.6. *There is only one infinite complete extension of the theory \mathcal{E} and that is the theory \mathcal{E}_∞ .*

PROOF. That \mathcal{E}_∞ is a complete extension of \mathcal{E} follows from Lemma 4.5. Let \mathcal{E}_* be any other infinite complete extension of \mathcal{E} . We have $\mathcal{E}_* \neq \mathcal{E}_m$ for all $m < \omega$. In the notation of Lemma 4.2, $\Delta_m \notin \mathcal{E}_*$ for all $m < \omega$. Hence $\neg \Delta_m \in \mathcal{E}_*$ for all $m < \omega$. This last implies in view of Lemma 4.4, that $\mathcal{E}_\infty \subseteq \mathcal{E}_*$. Since both these theories are complete, we conclude that $\mathcal{E}_\infty = \mathcal{E}_*$, as was to be shown.

THEOREM 4.7. *The theories \mathcal{E} and \mathcal{E}_∞ are decidable.*

PROOF. Let ϕ be any sentence. Count the number of variables in ϕ , say $m + 1$. Now $\phi \in \mathcal{E}_\infty$ if and only if $\phi \in \mathcal{E}_m$ by Theorem 3.4. Since the condition $\phi \in \mathcal{E}_m$ can be decided effectively, we have an effective decision procedure for \mathcal{E}_∞ . Finally, notice that $\phi \in \mathcal{E}$ if and only if $\phi \in \bigcap_{n < m} \mathcal{E}_n$; again a condition that can be checked in a finite number of steps. The proof is complete.

THEOREM 4.8. *For any formula ϕ with all free variables in the set $\{v_0, \dots, v_{p-1}\}$, it can be decided effectively whether ϕ defines a p -ary geometrical relation in terms of the geometrical relations B and E .*

PROOF. The formula ϕ will contain only $m + 1$ variables. According to Theorem 3.6, we need only check whether ϕ defines a geometrical relation with respect to Euclidean spaces of at most dimension m . In fact, it is sufficient to restrict attention to one Euclidean space \mathfrak{B} of dimension exactly m and only consider the identity isometries from Euclidean subspaces of \mathfrak{B} onto themselves. This checking can be carried out by seeing if the relation defined by ϕ and restricted to a subspace is the same relation obtained by restricting all the free variables in ϕ to the subspace and relativising all the quantifiers in ϕ to the subspace. But, the predicate of being in the least subspace spanned by a given number of points is definable in first-order logic in terms of betweenness and equidistance. Thus, since the number of points needed for specifying a subspace is at most $m + 1$, we can translate the question of the equivalence of the two forms of the relation defined by ϕ into a single first-order sentence. This sentence, then, need only be checked for validity in dimension m , a process that is effective.

This completes the formal development of the subject, and the author would like to conclude with some informal remarks. An amusing point to notice in the arguments of this section is that any sentence Δ_m satisfying the conditions of Lemma 4.2 must necessarily contain at least $m + 2$ variables. That the lower bound can actually be attained in the theory of B and E can be verified by writing out in logical symbols the sentence given in words in the proof of Lemma 4.1.

The consequence of these results for the problem of axiomatizing Euclidean geometry is that the theory \mathcal{E} is the only one of these theories that need be axiomatized, for we have shown above that one may pass from \mathcal{E} to \mathcal{E}_m simply by the adjunction of the sentences Δ_m . Though all details have not been completely checked by the author, it would seem that an adequate axiomatization of the theory \mathcal{E} would result by dropping axioms A11 and A12 of the system given by Tarski in [3]. Finally, the simplest way of axiomatizing infinite dimensional geometry would be to add to \mathcal{E} an infinite list of sentences expressing the fact that any number of mutually equidistant points can be found.

This last remark about infinite dimensional geometry indicates an immediate difference between the first-order formalism and theories permitting quantification over arbitrary finite sets as explained for the theory \mathcal{E}_2' in [3]. For it is seen at once that the infinite dimensional character of a space can be expressed in a *single* sentence involving

variables ranging over arbitrary finite sets, a fact clearly not true in the first-order theories in view of Lemma 4.5. Further, there seems to be no hope of giving a simple syntactical method like the counting of the number of variables for showing the relation of the truth of a sentence in one dimension to the truth in another dimension in these extended theories as was done in the fundamental result for our investigation Theorem 3.4. However, Tarski has noticed that Corollary 3.5 about infinite dimensional Euclidean spaces still holds for properties of relations formulated in the extended theory with finite sets, because the result in Theorem 3.1 of [2] remains valid in this generalization. Hence, even from this broader view, there is no way to distinguish between infinite dimensional Euclidean spaces.

Bibliography

- [1] TARSKI, A., *A decision method for elementary algebra and geometry*. Second edition, Berkeley and Los Angeles 1951, VI+63 pp.
- [2] ——— and VAUGHT, R. L., *Arithmetical extensions of relational systems*. *Compositio Mathematica*, vol. 13 (1957), pp. 81–102.
- [3] ——— *What is elementary geometry?* This volume, pp. 16–29.

BINARY RELATIONS AS PRIMITIVE NOTIONS IN ELEMENTARY GEOMETRY

RAPHAEL M. ROBINSON

University of California, Berkeley, California, U.S.A.

1. Introduction. We shall consider *equidistance* and the *order* of points on a line as the standard primitive notions of Euclidean, hyperbolic, or elliptic geometry. Here equidistance is a quaternary relation, whereas the order of points on a line is described in Euclidean or hyperbolic geometry by the ternary relation of *betweenness*, and in elliptic geometry by the quaternary relation of *cyclic order*. Various axiom systems have been given in terms of these primitive notions; see, for example, Tarski [7] for the Euclidean case. The adequacy of other proposed primitive notions for geometry will be judged by comparison with the standard ones.

M. Pieri [4] has shown that a *ternary* relation, that of a point being equally distant from two other points, can be used as the only primitive notion of Euclidean geometry of two or more dimensions. Indeed, in terms of this relation, it is possible to define equidistance of points in general, and the order of points on a line. The same ternary relation is also a possible primitive notion for either of the non-Euclidean geometries, hyperbolic or elliptic. A detailed discussion of Pieri's relation is given in Section 2.

We may raise the question whether one or more *binary* relations might serve as the primitive notions in some of the geometries. This is impossible in Euclidean geometry as described above, since the primitive notions of equidistance and order are preserved by similarity transformations, and no non-trivial binary relation is so preserved. However, let us choose a *unit distance* in the Euclidean space, and regard the property of two points being a unit distance apart as a new primitive notion. Then only isometric transformations preserve the primitive notions. The problem concerning the possibility of using just binary relations as primitive notions is thus reinstated.

We shall suppose at all times that a p -dimensional Euclidean, hyperbolic, or elliptic space is under discussion.¹ The exact value of p is usually immaterial, except that *we shall always suppose that $p \geq 2$* ;

¹ Many of the results stated for elliptic geometry apply also to spherical geometry, but we shall not go into this.

this condition will be understood henceforth without explicit mention. (The one-dimensional case will be excluded because the results there are usually exceptional and rather trivial.) Only the standard case where the base field is the field of real numbers will be considered. Thus, for example, the Euclidean p -space will be regarded as the direct p th power of the field of real numbers. The points of the space will be denoted by A, B, C, \dots, X, Y, Z , or by these letters with subscripts or superscripts. (In contrast to this, the letters a, b, c, \dots, x, y, z will be used for real numbers, with i, j, k, \dots, p, q, r reserved for natural numbers).

The space will be regarded as a metric space, the distance from A to B being denoted by AB . Thus the symbol AB always denotes a non-negative real number (unless it occurs as part of a formula, such as $AB \perp CD$, which is defined as a whole). In the Euclidean case, the distance AB is the square root of the sum of the squares of the differences of the coordinates of A and B . In the non-Euclidean cases, the metric will be assumed to be chosen so that the natural unit of length is being used.

*The definability of a notion in terms of given notions will always be understood in this paper as elementary (= arithmetical) definability.*² That is, aside from the given notions, a definition will use only the concepts of elementary logic, and the only variables used will be A, B, C, \dots , which range over points of the given space. The following logical symbols will be used: \wedge (and), \vee (or), \neg (not), \rightarrow (if \dots then \dots), \leftrightarrow (if and only if), Λ (for every), and \mathbf{V} (there exists). Identity (between points of the given space) will also be regarded as a logical concept. In addition, we shall sometimes use equations such as $AB = CD$. Here the entire equation may be regarded as a convenient notation for the quaternary relation of equidistance between points.

We return now to the question whether some of the geometries might be based on primitive notions which are all binary relations. In other words, *are there some binary relations which are definable in terms of the usual primitive notions, and in terms of which the usual primitive notions are definable?* We shall show that in elliptic geometry, it is possible to use a single binary relation as the only primitive notion.³ In particular,

² Some problems concerning more general types of definability are studied by Royden [5].

³ This result was found independently by H. L. Royden and the author, shortly after listening to a lecture by Alfred Tarski on the primitive notions of Euclidean geometry, based in part on Beth and Tarski [1], in the spring of 1956. The binary relation used by Royden was $AB \leq \pi/4$.

as shown in Section 3, the binary relation $AB = \pi/2$, which expresses that the two points A and B are at a distance $\pi/2$ apart (which is the maximum possible distance in the elliptic space), is a suitable primitive notion for elliptic geometry.

On the other hand, in Euclidean geometry (with a unit distance given), or in hyperbolic geometry, it is impossible to use binary relations as the only primitive notions. This is proved in Section 4 for binary relations of the form $AB = d$, and in Section 5 for binary relations in general. The difference between these geometries and elliptic geometry is due mainly to the fact that the elliptic space is bounded. In fact, as shown in Section 6, the local properties of Euclidean and hyperbolic spaces are expressible in terms of a binary relation, that of two points being at a prescribed distance apart.

If d is chosen so that the distance d is definable in terms of the usual primitive notions, then the system based on the binary relation $AB = d$ as its only primitive notion is weaker than the standard one so far as definability of concepts is concerned, but otherwise it is incomparable. The distance d cannot always be definable, since there are a non-denumerable infinity of possible values of d , but only a denumerable infinity of possible definitions. The problem of determining which distances d are definable is solved in Section 7.

2. Pieri's ternary relation. As mentioned in Section 1, Pieri has shown that in Euclidean geometry, it is possible to define the equidistance relation $AB = CD$ and betweenness in terms of the ternary relation $AB = BC$. His argument is also valid in hyperbolic geometry. We give below a proof somewhat different than Pieri's, and then show how to extend it to the elliptic case.

THEOREM 2.1. *Pieri's ternary relation $AB = BC$ is a suitable primitive notion for Euclidean, hyperbolic, or elliptic geometry.*

PROOF. Let the symbols $\text{bet}(A, B, C)$, $\text{col}(A, B, C)$, and $\text{sym}(A, B, C)$ express respectively that B is between A and C , that A, B, C are collinear, and that A and C are symmetric with respect to B (that is, that B is the midpoint of the segment joining A and C). Then the following definitions are valid formulas in Euclidean or hyperbolic geometry:

$$AB \leq BC \leftrightarrow (\forall X)[BX = XC \rightarrow (\forall Y)(AY = YB = BX)],$$

$$\text{bet}(A, B, C) \leftrightarrow B \neq A \wedge B \neq C \wedge (\bigwedge X)[XA \leq AB \wedge XC \leq CB \rightarrow X = B],$$

$$\text{col}(A, B, C) \leftrightarrow A = B \vee A = C \vee B = C$$

$$\vee \text{bet}(B, A, C) \vee \text{bet}(A, B, C) \vee \text{bet}(A, C, B),$$

$$\text{sym}(A, B, C) \leftrightarrow (\bigwedge X)[\text{col}(A, B, X) \wedge AB = BX \leftrightarrow X = A \vee X = C],$$

$$AB = CD \leftrightarrow (\bigvee X, Y)[\text{sym}(A, X, C) \wedge \text{sym}(B, X, Y) \wedge YC = CD].$$

Hence betweenness and equidistance are definable in terms of Pieri's relation, as was to be shown.

We shall now extend the result to the elliptic case.⁴ Some modifications of the above definitions are required. The definition of $AB \leq BC$ is still correct. The validity of the next definition depends on how we interpret $\text{bet}(A, B, C)$ in elliptic geometry. It is correct if we understand this to mean that there is a unique shortest line segment joining A and C , and that B is an interior point of this segment. Notice that when $AC = \pi/2$, as well as when $A = C$, there is no such point B .

The definition of $\text{col}(A, B, C)$ given above is not valid in elliptic geometry. We shall give a definition below which expresses collinearity as a special case of cyclic order, which we also need. Once collinearity has been defined, the previous definitions of $\text{sym}(A, B, C)$ and $AB = CD$ may be used. Notice that the definition of $\text{sym}(A, B, C)$ is so formulated that the relation holds, as it should, when $A = C$ and $AB = \pi/2$.

We now wish to define the cyclic order of points on a line. We start by defining recursively a relation $\text{seq}(A_0, A_1, \dots, A_n)$ for $n \geq 2$, as follows:

$$\text{seq}(A_0, A_1, A_2) \leftrightarrow \text{bet}(A_0, A_1, A_2),$$

$$\text{seq}(A_0, A_1, \dots, A_n) \leftrightarrow \text{seq}(A_0, A_1, \dots, A_{n-1})$$

$$\wedge \text{bet}(A_{n-2}, A_{n-1}, A_n) \wedge A_n \neq A_0 \wedge \neg \text{bet}(A_{n-1}, A_0, A_n).$$

It is seen that $\text{seq}(A_0, A_1, \dots, A_n)$ expresses that the sequence of points A_0, A_1, \dots, A_n lie in this cyclic order on a line, and divide the line into intervals such that, excluding the one from A_n to A_0 , the sum of the lengths of any two consecutive intervals is less than $\pi/2$. This extraneous condition concerning the lengths of the intervals may be removed by

⁴ A reader who is concerned only with Euclidean and hyperbolic geometry may proceed directly to Section 4.

putting

$$\text{ord}(A_0, A_1, \dots, A_n) \leftrightarrow (\forall X_0, X_1, \dots, X_{4n})[X_0 = A_0 \wedge X_4 = A_1 \\ \wedge \dots \wedge X_{4n} = A_n \wedge \text{seq}(X_0, X_1, \dots, X_{4n})].$$

Then, as is easily seen, $\text{ord}(A_0, A_1, \dots, A_n)$ expresses simply that the points A_0, A_1, \dots, A_n are in this cyclic order on a line. In particular, $\text{ord}(A, B, C, D)$ is the basic quaternary relation of cyclic order in elliptic geometry. Furthermore,

$$\text{col}(A, B, C) \leftrightarrow A = B \vee A = C \vee B = C \vee \text{ord}(A, B, C),$$

so that, as previously noted, equidistance is also definable. Thus Pieri's ternary relation is a suitable primitive notion for elliptic geometry. (It may be noticed that the relation $\text{seq}(A_0, A_1, \dots, A_n)$, defined above for all $n \geq 2$, was needed only for $n \leq 12$.)

3. Binary primitives for elliptic geometry. In elliptic geometry, Pieri's relation is definable in terms of any distance d with $0 < d \leq \pi/2$. The converse holds only if $\cos d$ is algebraic. In this case, the binary relation $AB = d$ is a suitable primitive notion for elliptic geometry. A detailed proof is given only for $d = \pi/2$, which seems to be the most interesting case, since the condition $AB = \pi/2$ expresses that the polar of either of the points A or B passes through the other.

We start by noticing two definitions that can be used in the elliptic plane. The formulas

$$\text{col}(A, B, C) \leftrightarrow (\forall X)[AX = BX = CX = \pi/2]$$

and

$$AB \perp CD \leftrightarrow A \neq B \wedge C \neq D \wedge (\forall X)[AX = BX = \pi/2 \wedge \text{col}(C, D, X)]$$

define the collinearity of three points and the perpendicularity of two lines, in terms of the distance $\pi/2$. Here, of course, a notation such as $AX = BX = \pi/2$ is short for $AX = \pi/2 \wedge BX = \pi/2$; the concept of equidistance is not involved.

THEOREM 3.1. *The following formula holds in the elliptic plane:*

$$BC = CA = AB = \pi/2 \wedge \text{col}(B, P, C) \wedge \text{col}(C, Q, A) \wedge \text{col}(A, R, B) \\ \wedge P \neq C \wedge Q \neq C \wedge AP \perp QR \wedge BQ \perp PR \rightarrow AR = \pi/4.$$

PROOF. One model of the elliptic plane consists of all lines through the origin in a three-dimensional Euclidean space. We may identify A , B , C with the x , y , z axes. Then AP and BQ correspond to planes $z = ay$ and $z = bx$, with suitable values of a and b . The planes through P perpendicular to BQ and through Q perpendicular to AP are

$$x - aby + bz = 0, \quad -abx + y + az = 0.$$

These planes intersect the plane $z = 0$ in lines where $x = aby$ and $y = abx$, respectively. These lines coincide only if $ab = \pm 1$. Thus the point R is represented by one of the lines $y = \pm x$, $z = 0$.

THEOREM 3.2. *The binary relation $AB = \pi/2$ is a suitable primitive notion for elliptic geometry.*

PROOF. We can define the distance $\pi/2$ in terms of Pieri's relation, by using the definition of $\text{bet}(A, B, C)$ from Section 2, and the formula

$$AB = \pi/2 \leftrightarrow A \neq B \wedge \neg(\forall X) \text{bet}(A, X, B).$$

It remains to define Pieri's relation in terms of the distance $\pi/2$.

Consider first elliptic plane geometry. Notice that the distance $\pi/4$ is definable. Indeed, we see that $AR = \pi/4$ if and only if there exist points B, C, P, Q satisfying the conditions stated in Theorem 3.1.

We now give a series of further definitions leading to Pieri's relation $AB = BC$:

$$\text{mid}(A, B, C) \leftrightarrow \text{col}(A, B, C) \wedge (\forall X)[AX = CX = \pi/4 \wedge AC \perp BX],$$

$$\text{mex}(A, B, C) \leftrightarrow \text{col}(A, B, C) \wedge (\forall X)[\text{mid}(A, X, C) \wedge BX = \pi/2],$$

$$\text{sym}(A, B, C) \leftrightarrow \text{mid}(A, B, C) \vee \text{mex}(A, B, C) \vee A = B = C$$

$$\vee (A = C \wedge AB = \pi/2) \vee (AC = \pi/2 \wedge AB = BC = \pi/4),$$

$$AB = BC \leftrightarrow (\forall X)[\text{sym}(A, X, C) \wedge BX = \pi/2].$$

Here the conditions $\text{mid}(A, B, C)$ and $\text{mex}(A, B, C)$ require that $A \neq C$ and $AC \neq \pi/2$, and that B is the midpoint of the shorter or longer line segment joining A and C (the "internal" or "external" midpoint). The definition of $\text{sym}(A, B, C)$ then gives a complete listing of the cases in which A and C are symmetric with respect to B . Notice that, in the definition of $AB = BC$, if $A \neq C$, then there are just two possible values of X , the midpoints of the two segments joining A and C , and the polar of either is perpendicular to AC at the other. If $A = C$, then X may be A or

any point on the polar of A , and B is completely arbitrary, as it should be.

The restriction to plane geometry may be removed by noticing that it is possible to define the concept of a plane in p -space in terms of the distance $\pi/2$. We can then define the relation $AB = BC$ by applying the previous method in a plane containing A , B , and C .

THEOREM 3.3. *In elliptic geometry, equidistance is definable in terms of the distance d , for any d with $0 < d \leq \pi/2$.*

This can be derived from Theorem 3.2, by defining the distance $\pi/2$ in terms of the distance d , but the details of the proof will be omitted. Combining this result with Theorem 7.3, we see that the binary relation $AB = d$ is a suitable primitive notion for elliptic geometry if and only if $0 < d \leq \pi/2$ and $\cos d$ is algebraic.

4. Patch-wise congruence. Let any Euclidean or hyperbolic space be given. Then we put

$$\text{con}(X_1, X_2, \dots, X_m; X_1', X_2', \dots, X_m') \leftrightarrow$$

$$X_1X_2 = X_1'X_2' \wedge X_1X_3 = X_1'X_3' \wedge \dots \wedge X_{m-1}X_m = X_{m-1}'X_m'.$$

That is, two finite sequences of points are called congruent if all the corresponding distances are equal. The space has a certain property of homogeneity expressed by the condition

$$\text{con}(X_1, \dots, X_m; X_1', \dots, X_m') \rightarrow$$

$$(\wedge Y)(\vee Y') \text{con}(X_1, \dots, X_m, Y; X_1', \dots, X_m', Y'),$$

which holds for all values of m . The only other fact about the given space that we use in this section is that the space is unbounded.

The concept of congruence will now be extended to that of *patch-wise congruence*. If $c > 0$, the formula

$$\text{pat}(c: X_1, X_2, \dots, X_m; X_1', X_2', \dots, X_m')$$

will be used to denote that the two sequences X_1, X_2, \dots, X_m and X_1', X_2', \dots, X_m' are patch-wise congruent, with separation constant c . This formula is defined as follows. We start by considering any partition of the indices $1, 2, \dots, m$. For this partition, we form the conjunction of all the formulas $X_iX_j = X_i'X_j'$ for i and j in the same class, and of all the formulas $X_iX_j > c$ and $X_i'X_j' > c$ for i and j in different classes. The disjunction of all these conjunctions, formed for all possible partitions, is the required formula expressing patch-wise congruence.

The formula $\text{pat}(c: X_1, \dots, X_m; X'_1, \dots, X'_m)$ constructed in this way actually expresses that the two sequences of points can be divided into patches which are respectively congruent, such that the distance between any two patches is greater than c . We shall now show that the formula expressing the property of homogeneity mentioned above may be extended to patch-wise congruence.

THEOREM 4.1. *In any Euclidean or hyperbolic space, and for any $c > 0$, we have*

$$\text{pat}(2c: X_1, \dots, X_m; X'_1, \dots, X'_m) \rightarrow (\wedge Y)(\vee Y') \text{pat}(c: X_1, \dots, X_m, Y; X'_1, \dots, X'_m, Y').$$

PROOF. Pick out one disjunct of the hypothesis which is valid. This determines which points are to be considered as belonging to the same patch. Now if Y is at a distance greater than c from all the points X_i , then it may be considered as forming a new patch, and we may choose for Y' any point at a distance greater than c from all the points X'_i . Otherwise, there is a unique patch, among the points X_1, \dots, X_m , such that Y is at a distance at most c from some point of the patch. Choose Y' in a corresponding position relative to the corresponding patch of the points X'_i .

THEOREM 4.2. *Let d be a positive number. Let a Euclidean or hyperbolic space be given. Let $\alpha_1, \alpha_2, \dots, \alpha_q$ be binary relations on this space, such that for $k = 1, 2, \dots, q$, we have*

$$XY = X'Y' \rightarrow [\alpha_k(X, Y) \leftrightarrow \alpha_k(X', Y')]$$

and

$$\alpha_k(X, Y) \rightarrow XY \leq d.$$

Let ϕ be a formula with free variables X_1, \dots, X_m , which is elementary in terms of $\alpha_1, \dots, \alpha_q$; that is, the atomic formulas of ϕ have the form $X = Y$ or $\alpha_k(X, Y)$. Then

$$\text{pat}(2^nd: X_1, \dots, X_m; X'_1, \dots, X'_m) \rightarrow [\phi(X_1, \dots, X_m) \leftrightarrow \phi(X'_1, \dots, X'_m)],$$

*provided that ϕ does not contain more than n nested quantifiers.*⁵

⁵ It can be shown that 2^nd is the smallest possible separation constant which can be used here.

PROOF. By induction in n . The result is clear for $n = 0$, since on the basis of the hypothesis about patch-wise congruence, we have, for any possible i and j , either $X_i X_j = X_i' X_j'$, or else both $X_i X_j > d$ and $X_i' X_j' > d$. Hence

$$X_i = X_j \leftrightarrow X_i' = X_j', \quad \alpha_k(X_i, X_j) \leftrightarrow \alpha_k(X_i', X_j')$$

for all values of k , and the conclusion follows.

We now assume the theorem for some value of n , and prove it for $n + 1$. It will be sufficient to consider the case in which

$$\phi(X_1, \dots, X_m) \leftrightarrow (\forall Y) \psi(X_1, \dots, X_m, Y),$$

where ψ is an elementary formula with the indicated free variables and containing at most n nested quantifiers. (For the truth-value of any admissible formula ϕ can be determined from the truth-values of formulas of this form.) Now by the inductive hypothesis, we have

$$\text{pat } (2^nd: X_1, \dots, X_m, Y; X_1', \dots, X_m', Y') \rightarrow$$

$$[\psi(X_1, \dots, X_m, Y) \leftrightarrow \psi(X_1', \dots, X_m', Y')],$$

and according to Theorem 4.1,

$$\text{pat } (2^{n+1}d: X_1, \dots, X_m; X_1', \dots, X_m') \rightarrow$$

$$(\Lambda Y)(\forall Y') \text{ pat } (2^nd: X_1, \dots, X_m, Y; X_1', \dots, X_m', Y').$$

Combining these results, we see that

$$\text{pat } (2^{n+1}d: X_1, \dots, X_m; X_1', \dots, X_m') \rightarrow$$

$$[(\forall Y) \psi(X_1, \dots, X_m, Y) \leftrightarrow (\forall Y') \psi(X_1', \dots, X_m', Y')].$$

THEOREM 4.3. *Under the same hypotheses on $\alpha_1, \alpha_2, \dots, \alpha_q$, equidistance is not definable in terms of them.*

PROOF. Suppose the relation $X_1 X_2 = X_3 X_4$ were definable in terms of $\alpha_1, \dots, \alpha_q$, using a formula containing at most n nested quantifiers. Then, by Theorem 4.2, we would have

$$\text{pat } (2^nd: X_1, X_2, X_3, X_4; X_1', X_2', X_3', X_4') \rightarrow$$

$$[X_1 X_2 = X_3 X_4 \leftrightarrow X_1' X_2' = X_3' X_4'].$$

This is certainly false, since the hypothesis holds whenever we have $X_i X_j > 2^nd$ and $X_i' X_j' > 2^nd$ for all i and j .

THEOREM 4.4. *In Euclidean or hyperbolic geometry, equidistance is not definable in terms of any number of particular distances.*

PROOF. For $k = 1, 2, \dots, q$, let $\alpha_k(X, Y) \leftrightarrow XY = d_k$, where $d_k > 0$. Then $\alpha_1, \alpha_2, \dots, \alpha_q$ satisfy the hypotheses of Theorem 4.2, if we take $d = \max(d_1, d_2, \dots, d_q)$. Now apply Theorem 4.3.

5. No binary primitives for Euclidean or hyperbolic geometry. We now come to the question whether there are any binary relations $\alpha_1(X, Y), \dots, \alpha_q(X, Y)$ which are suitable primitive notions for Euclidean geometry (with a unit distance) or for hyperbolic geometry. To be suitable, they should be definable in terms of equidistance and the unit distance in the Euclidean case, and in terms of equidistance alone in the hyperbolic case, and conversely. To show that this is impossible, we start by studying binary relations which are definable in terms of equidistance and r particular distances d_1, d_2, \dots, d_r . (Actually, we need only $r = 1, d_1 = 1$ in the Euclidean case, and $r = 0$ in the hyperbolic case.) In the hyperbolic case, a preliminary theorem is needed.

THEOREM 5.1. *In hyperbolic p -space, it is possible to introduce coordinates (x_1, x_2, \dots, x_p) , with $x_1^2 + x_2^2 + \dots + x_p^2 < 1$, so that e^{AB} can be calculated from the coordinates of A and B using rational operations and the extraction of square roots.*

PROOF. In the interior of the unit sphere in Euclidean p -space introduce a new metric $[A, B]$ by putting $[A, B] = 0$ if $A = B$, and

$$[A, B] = \frac{1}{2} \left| \log \frac{AR \cdot BS}{BR \cdot AS} \right|$$

otherwise, where R and S are the two points where the line joining A and B intersects the unit sphere. From the coordinates of A and B , we can calculate successively the coordinates of R and S , the distances AR, BS, BR, AS , and finally $e^{[A, B]}$, using only rational operations and the extraction of square roots. Now it is known that, with the metric just introduced, the interior of the unit sphere in Euclidean p -space becomes a model for hyperbolic p -space. (See, for example, Hilbert and Cohn-Vossen [2], § 35.) Thus the theorem restates, from a different viewpoint, what we have just proved.

THEOREM 5.2. *In a Euclidean or hyperbolic space, any binary relation $\alpha(X, Y)$ which is definable in terms of equidistance and particular distances*

d_1, d_2, \dots, d_r , satisfies the condition

$$XY = X'Y' \rightarrow [\alpha(X, Y) \leftrightarrow \alpha(X', Y')],$$

and, for some $d > 0$, one of the conditions

$$XY > d \rightarrow \alpha(X, Y), \quad XY > d \rightarrow \neg\alpha(X, Y).$$

PROOF. The first conclusion is clear, since there is an isometric mapping of the space onto itself which takes X into X' and Y into Y' . This mapping preserves the equidistance relation and the particular distances, and hence anything definable in terms of them.

We turn now to the second conclusion. Suppose that $t > 0$, and consider the formula

$$(\wedge X, Y)[XY = t \rightarrow \alpha(X, Y)].$$

We can eliminate all point variables in favor of real variables, by introducing coordinates. In the Euclidean case, by simply squaring all equations that occur, we obtain an equivalent formula of elementary algebra, containing only t and d_1, d_2, \dots, d_r as free variables. In the hyperbolic case, we use Theorem 5.1; by a little manipulation, including the elimination of square roots by introducing additional existential quantifiers, we again obtain an equivalent formula of elementary algebra, where in this case e^t and e^{d_1}, \dots, e^{d_r} play the role of free variables.

Following the procedure of Tarski [6], all bound variables can be eliminated, if we allow the introduction of inequalities (Tarski's Theorem 31). If numerical values are assigned to d_1, d_2, \dots, d_r , we see that there is a real number d such that the resulting formula is either true for all $t > d$ or else false for all $t > d$. Thus the same alternatives hold for the displayed formula with which we started. In the first case, we have $XY > d \rightarrow \alpha(X, Y)$. In the second, taking account of the fact that the truth-value of $\alpha(X, Y)$ depends only on XY , we see that $XY > d \rightarrow \neg\alpha(X, Y)$.

THEOREM 5.3. *In a Euclidean or hyperbolic space, it is impossible to find binary relations $\alpha_1, \alpha_2, \dots, \alpha_q$, which are definable in terms of equidistance and particular distances, and in terms of which equidistance is definable. Thus there are no binary relations which are suitable as the primitive notions of Euclidean geometry (with a unit distance) or of hyperbolic geometry.*

PROOF. We may apply Theorem 5.2 to each of the relations α_k . By replacing α_k by $\neg\alpha_k$ if necessary, we may assume that we have $XY > d \rightarrow$

$\neg\alpha_k(X, Y)$, and hence $\alpha_k(X, Y) \rightarrow XY \leq d$, for all values of k . The proof is completed by applying Theorem 4.3.

6. Local definability of equidistance. Although, as shown in Section 4, equidistance is not definable in terms of particular distances in Euclidean or hyperbolic geometry, nevertheless equidistance is *locally* definable in terms of a single given distance.⁶ By a local definition of equidistance $AB = CD$ (in terms of a given distance d) is meant a formula which provides a necessary and sufficient condition for this equality, on the assumption that the distance between each two of the four points does not exceed a prescribed bound. We shall see that this bound can be taken arbitrarily large, although the formula required becomes longer as the bound increases. (Throughout this section, d denotes an arbitrary positive number.)

THEOREM 6.1. *In Euclidean or hyperbolic geometry:*

- (a) *The distance $2d$ is definable in terms of the distance d .*
- (b) *Any one of the relations $AB = d$, $AB \leq d$, $AB < d$ is definable in terms of any other one.*
- (c) *The local symmetry relation $\text{sym}(A, B, C) \wedge AB \leq h$ is definable in terms of the distance d if and only if the distance h is definable in terms of the distance d .*

PROOF. (a) We may use the formula⁷

$$AB = 2d \leftrightarrow (\forall X)(\wedge Y)[AY = d \wedge BY = d \leftrightarrow Y = X].$$

⁶ At the time this paper was presented to the Symposium, I knew this result only for Euclidean or hyperbolic geometry of three or more dimensions. A few days afterwards, A. Seidenberg pointed out to me that the linkage of Peaucellier, which enables one to draw a line segment in the Euclidean plane, furnishes a local definition of collinearity in terms of a particular distance, and that this in turn leads to a local definition of equidistance. Some time later, Seidenberg also succeeded in extending the result to the hyperbolic plane. Subsequently, the author found a different and simpler solution to this problem. The method used here can also be adapted to the higher-dimensional hyperbolic spaces, and is presented below in this extended form. The local definition of midpoint used in the proof of Theorem 6.1(c) is a modified form of the definition suggested by Seidenberg for use in the hyperbolic plane.

⁷ This definition of the distance $2d$ in terms of the distance d uses an existential and a universal quantifier. In Euclidean geometry, it is also possible to define the distance $2d$ *existentially* in terms of the distance d , that is, by means of a formula in prenex form containing only existential quantifiers. (In the two-dimensional case,

(b) Whichever of the three relations is given, we can easily define $AB < 2d$, since this expresses that the spheres of radius d about A and B overlap. If the given relation is $AB \leq d$, then we may use the formula

$$AB < d \leftrightarrow (\bigwedge X)[AX \leq d \rightarrow BX < 2d]$$

to define $AB < d$, and hence $AB = d$ can also be defined. A similar argument applies in the other cases.

(c) We see that

$$0 < AB < 2d \rightarrow [\text{sym}(A, B, C) \leftrightarrow B \neq C$$

$$\wedge (\bigwedge X, Y)(AX = XB = BY = d \wedge XY = 2d \rightarrow CY = d)].$$

Indeed, the possible values of X lie on the intersection of two spheres $AX = d$ and $BX = d$, and, since $0 < AB < 2d$, these spheres actually intersect. The possible values of Y are those symmetric to X with respect to B . The only point C , other than B itself, which is at a distance d from all such points Y is the point symmetric to A with respect to B . This formula clearly leads to a suitable definition of the relation $\text{sym}(A, B, C) \wedge AB \leq d$, and the stated result then follows easily.

THEOREM 6.2. *In Euclidean or hyperbolic geometry, the local Pieri relation $AB = BC \leq h$ can be defined in terms of the distance d if and only if the distance h can be defined in terms of the distance d .*

PROOF. The necessity of the condition follows from Theorem 6.1(b). To prove the sufficiency, we need only show that the relation $AB = BC \leq d$ is definable in terms of the distance d . The proof is divided into three cases.

CASE 1. Euclidean p -space, $p \geq 3$. It is easily seen that

$$AB = BC \leq 2d \leftrightarrow (\bigvee X, Y, Z)[AY = YC = 2d$$

$$\wedge AX = XY = YZ = ZC = XB = BZ = d].$$

This is based on the idea of taking an isosceles triangle whose equal sides are $2d$ and folding it along the line joining the midpoints of the two equal

use a network of equilateral triangles. In three dimensions, use twice the fact that the diagonal of an octahedron is $2^{\frac{1}{2}}$ times an edge, and similarly in higher dimensions.) Starting from this fact, it is possible to put the local definition of equidistance in an existential form, and to define all algebraic distances existentially in terms of the unit distance (thus sharpening Theorems 6.3 and 7.1). I do not see any way of doing the corresponding things in the non-Euclidean cases.

sides. The vertex remains at an equal distance from the two ends of the base, and this distance may be any amount not less than half the base and not exceeding $2d$. By adjoining the condition $AB \leq d$, we obtain the required relation $AB = BC \leq d$.

CASE 2. Euclidean plane. There is a well-known linkage, due to Peaucellier, which can be used to draw a line segment. (See Kempe [3] or Hilbert and Cohn-Vossen [2], § 40.) Choosing, for the lengths of all links, distances definable in terms of d (for example, suitable multiples of d), and considering three positions of the linkage, we obtain a local definition of collinearity in terms of the distance d . Combining that with the formula

$$AB = BC \wedge AC < 2d \leftrightarrow$$

$$(\vee X, Y)[X \neq Y \wedge AX = CX = AY = CY = d \wedge \text{col}(B, X, Y)],$$

we easily obtain the required result.

CASE 3. Hyperbolic p -space, $p \geq 2$. If $p \geq 3$, we could proceed much as in Case 1. However, we shall apply a different method, which does not exclude the case $p = 2$, but which definitely uses the non-Euclidean character of the space. In fact, we see that

$$AC = 2d \rightarrow \{\text{col}(A, B, C) \leftrightarrow$$

$$(\vee X, Y)[\text{sym}(A, X, B) \wedge \text{sym}(B, Y, C) \wedge XY = d]\}.$$

We have expressed the similarity of the triangles ABC and XPY , which is impossible unless the triangles are degenerate. Indeed, in hyperbolic geometry, the line joining the midpoints of two sides of a triangle is less than half as long as the third side. Since $\text{sym}(A, B, C)$ is locally definable, we can obtain a local definition of $\text{col}(A, B, C)$, at least under the restriction that $AC = 2d$. From this, we can get a local definition of $\text{col}(B_1, B_2, B_3)$, without such a restriction, by considering three values of B with the same A and C . We can then proceed to a local definition of Pieri's relation as in Case 2.

THEOREM 6.3. *In Euclidean or hyperbolic geometry, the local equidistance relation*

$$AB = CD \wedge AB \leq h \wedge AC \leq h \wedge AD \leq h \wedge BC \leq h \wedge BD \leq h \wedge CD \leq h$$

can be defined in terms of the distance d if and only if the distance h is definable in terms of the distance d .

PROOF. The condition is clearly necessary, and the sufficiency can be derived from Theorem 6.2 by a suitable modification of the method used in Section 2.

7. Definable distances. We shall now determine what distances t are definable in terms of a given distance d , with or without the use of equidistance, or, in the non-Euclidean cases, in terms of equidistance alone.

We start by giving a few definitions valid in both the Euclidean and hyperbolic geometries. (With some modifications, they can be used also in the elliptic case.) The relation of equidistance is considered as given, and notions previously defined in terms of equidistance are also used. In the first place, we have

$$AB = CD + EF \leftrightarrow (AB = CD \wedge E = F) \vee (AB = EF \wedge C = D) \\ \vee (\forall X)[\text{bet}(A, X, B) \wedge AX = CD \wedge XB = EF].$$

We also wish to define perpendicularity. A special case is covered by the formula

$$AC \perp BC \leftrightarrow A \neq C \wedge B \neq C \wedge (\forall X)[\text{sym}(B, C, X) \wedge AB = AX].$$

We can then proceed to the formula

$$AB \perp CD \leftrightarrow A \neq B \wedge C \neq D \wedge (\forall X, Y, Z)[XZ \perp YZ \wedge \text{col}(A, X, Z) \\ \wedge \text{col}(B, X, Z) \wedge \text{col}(C, Y, Z) \wedge \text{col}(D, Y, Z)],$$

which defines perpendicularity in general.

THEOREM 7.1. *In Euclidean geometry, the distance t is definable in terms of equidistance and the unit distance if and only if t is algebraic. The algebraic distances are indeed definable in terms of the unit distance alone.*

PROOF. Suppose that $(\wedge A, B)[AB = t \leftrightarrow \phi(A, B)]$ is a valid formula of p -dimensional Euclidean geometry, where $\phi(A, B)$ is expressed in terms of equidistance and the unit distance. By introducing coordinates, it can be transformed into a formula of elementary algebra, with t as its only free variable. By Tarski [6], the bound variables may be eliminated, which leads to the conclusion that t must be algebraic.

It remains to show that all algebraic distances can be defined. We have already defined $AB = CD + EF$, and we can define the product of two

distances by the formula

$$AB = CD \cdot EF \leftrightarrow (A = B \wedge C = D) \\ \vee (\forall P, Q, R, S)[\text{col}(P, A, B) \wedge \text{col}(P, Q, R) \wedge P \neq A \\ \wedge AQ \perp PS \wedge BR \perp PS \wedge PQ = 1 \wedge PA = CD \wedge QR = EF].$$

Using these definitions of sum and product of two distances, we can express that a certain distance satisfies a given algebraic equation. By the use of suitable inequalities, which are also definable, we can isolate a particular root, and hence define that $AB = t$, where t is a given algebraic number.

If we are given only the unit distance, but not equidistance, then equidistance is nevertheless locally definable. All of the concepts used can be defined locally, which is sufficient for the purposes of the proof. (Notice that in the definition of the product of two distances above, we expressed the parallelism of the lines AQ and BR by the existence of a common perpendicular, and not by the non-existence of a point of intersection, so that this transition would be possible.)

THEOREM 7.2. *In hyperbolic geometry, the distance t is definable in terms of equidistance if and only if e^t is algebraic.*

PROOF. Using Theorem 5.1 and Tarski [6], we see that only such distances can be definable in terms of equidistance. It remains to show that all such distances are definable.

We have defined the relation $AB = CD + EF$, but the definition of $AB = CD \cdot EF$ does not apply here. Indeed, this product formula is not definable, since if it were, we could define the unit distance $AB = 1$, which is impossible since e is not algebraic. But we shall show that it is possible to define the two formulas

$$\cosh AB = \cosh CD + \cosh EF, \quad \cosh AB = \cosh CD \cdot \cosh EF.$$

We will then be able to express the condition that $\cosh AB$ satisfies a given algebraic equation, and hence the condition that it is a given algebraic number. Thus a distance t will be definable if $\cosh t$ is algebraic, or, what is equivalent, if e^t is algebraic.

The definition of the second formula follows at once from the known formula $\cosh c = \cosh a \cosh b$ connecting the sides of a right triangle. Thus we have

$$\cosh AB = \cosh CD \cdot \cosh EF \leftrightarrow (AB = CD \wedge E = F) \vee (AB = EF \wedge C = D) \\ \vee (\forall X)[AX \perp BX \wedge AX = CD \wedge BX = EF].$$

Also, since $2 \cosh x \cosh y = \cosh (x + y) + \cosh (x - y)$, we see that $2 \cosh AB = \cosh CD + \cosh EF \wedge CD \geq EF \leftrightarrow$

$$(\forall P, Q, R, S)[\cosh AB = \cosh PQ \cdot \cosh RS \\ \wedge CD = PQ + RS \wedge EF + RS = PQ],$$

which leads to a definition of $2 \cosh AB = \cosh CD + \cosh EF$. The factor 2 on the left could be removed, if we were able to define the relation $\cosh XY = 2$. This can be done, for example, by a judicious combination of the above formulas. Indeed, we see that

$$\cosh AB = 2 \leftrightarrow A \neq B \wedge (\forall P, Q, R, S)[\cosh PQ = \cosh^2 AB \\ \wedge 2 \cosh AB = \cosh RS + 1 \wedge 2 \cosh RS = \cosh AB + \cosh PQ].$$

Since $\cosh XX = 1$, we see that all the equations on the right are special cases of the formulas which we have defined, so that this furnishes the desired definition.

The proofs of the last two theorems will be omitted, since they do not require any essentially new methods.

THEOREM 7.3. *In elliptic geometry, the distance t is definable in terms of equidistance if and only if $\cos t$ is algebraic.*

THEOREM 7.4. *The distance t is definable in terms of the distance d (where $d > 0$, and in the elliptic case also $d \leq \pi/2$) if and only if the stated condition is satisfied.*

- (a) *Euclidean case: t/d is algebraic.*
- (b) *Hyperbolic case: e^t is algebraic in terms of e^d .*
- (c) *Elliptic case: $\cos t$ is algebraic in terms of $\cos d$.*

These results are unchanged if the relation of equidistance is also considered as given.

Bibliography

- [1] BETH, E. W. and A. TARSKI, *Equilaterality as the only primitive notion of Euclidean geometry*. *Indagationes Mathematicae*, vol. 18 (1956), pp. 462-467.
- [2] HILBERT, D. and S. COHN-VOSSEN, *Anschauliche Geometrie*. Berlin 1932, viii+310 pp. [English translation: *Geometry and the imagination*. New York 1956, ix+357 pp.]
- [3] KEMPE, A. B., *How to draw a straight line; a lecture on linkages*. London 1877, vi+51 pp.

- [4] PIERI, M., *La geometria elementare istituita sulle nozioni di 'punto' e 'sfera'*. Memorie di Matematica e di Fisica della Società Italiana delle Scienze, ser. 3, vol. 15 (1908), pp. 345–450.
- [5] ROYDEN, H. L., *Remarks on primitive notions for elementary Euclidean and non-Euclidean plane geometry*. This volume, pp. 86–96.
- [6] TARSKI, A., *A decision method for elementary algebra and geometry*. Second edition, Berkeley and Los Angeles 1951, iv+63 pp.
- [7] —, *What is elementary geometry?* This volume, pp. 16–29.

REMARKS ON PRIMITIVE NOTIONS FOR ELEMENTARY EUCLIDEAN AND NON-EUCLIDEAN PLANE GEOMETRY

H. L. ROYDEN

Stanford University, Stanford, California, U.S.A.

Introduction. The purpose of the present paper is to explore some relationships between primitive notions in elementary plane geometry with a view to determining the possibility of defining certain notions in terms of others. All of our primitive notions are predicates whose arguments are the primitive elements (points or points and lines) and we say that a primitive F can be *defined* in terms of a primitive G relative to a deductive system \mathbf{S} if

$$(x, y, z, \dots)[F(x, y, z, \dots) \Leftrightarrow \Phi(x, y, z, \dots)]$$

is a theorem in \mathbf{S} where Φ is a sentential function involving only G and logical terms in its formation (cf. [10]).

Whether F is definable in terms of G depends not only on the deductive system \mathbf{S} , but also on the logical basis used and our results are sometimes different if we use only the restricted predicate calculus rather than a logic which contains the theory of sets. Definitions using only the restricted predicate calculus will be called *elementary* and the others *set-theoretic*. In the present paper all of our definitions are elementary except for part of Section 5 where there is some discussion of the possibility of definitions using variables ranging over finite sets of points.

We consider here both Euclidean and non-Euclidean geometry and use a set of axioms equivalent to Hilbert's without the axioms of completeness and of Archimedes. We shall sometimes supplement these with an axiom (P12) to the effect that any line through a point inside a circle has a point in common with the circle. One of my purposes here is to show the role played by this axiom in the definability of concepts in elementary geometry.

Theorem 1 shows that for Euclidean and elliptic geometry this axiom plays an essential role in the possibility of defining order in terms of collinearity. With regard to hyperbolic geometry the situation is markedly different and order can be defined in terms of collinearity independently

of this axiom. As Menger [3, 4] has pointed out, the whole of hyperbolic geometry can be built on the notion of collinearity. We use here the elegant definition of order given by Jenks [2], but our treatment of the definition of congruence differs somewhat from that of Menger and his students in that we first define orthogonality and use it in the definition of congruence.

1. The basic elementary geometries. Euclidean geometry. We shall consider two systems for elementary Euclidean plane geometry. The first is the system \mathcal{P} which uses the undefined primitives β and δ and consists of all consequences of the axioms P1–P12 listed below. Intuitively, $\beta(xyz)$ has the meaning “ x , y , and z are collinear and y is between x and z ,” while $\delta(xyzw)$ has the meaning “the segment xy is congruent to the segment zw .” In terms of these notions we define the notion of *collinearity*:

$$\lambda(xyz) =_{\text{df}} \beta(xyz) \vee \beta(yzx) \vee \beta(zxy);$$

and *parallelism*:

$$\pi(xyuv) =_{\text{df}} \sim(\exists t)[\lambda(xyt) \ \& \ \lambda(uvt)].$$

Thus $\pi(xyuv)$ states that (x, y) and (u, v) are pairs of distinct points lying on distinct parallel lines. Our axiom system for \mathcal{P} corresponds to Hilbert's axiom system, with the exclusion of the axioms of Archimedes and of completeness, and is equivalent to the Axioms A1–12 of Tarski [12]. In fact, our axioms are taken directly from Tarski's paper, except that our version P7 of Pasch's axiom is stronger than Tarski's A7 and together with the remaining axioms it implies Tarski's A12, which is accordingly omitted from our list.

AXIOMS FOR \mathcal{P}

- P1 $(x)(y)[\beta(xyx) \Rightarrow x = y]$
 P2 $(x)(y)(z)(u)[\beta(xyu) \ \& \ \beta(yzu) \Rightarrow \beta(xyz)]$
 P3 $(x)(y)(z)(u)[\beta(xyz) \ \& \ \beta(xyu) \ \& \ (x \neq y) \Rightarrow \beta(xzu) \vee \beta(xuz)]$
 P4 $(x)(y)\delta(xyyx)$
 P5 $(x)(y)(z)[\delta(xyzz) \Rightarrow (x = y)]$
 P6 $(x)(y)(z)(u)(v)(w)[\delta(xyzu) \ \& \ \delta(xyvw) \Rightarrow \delta(zuvw)]$
 P7 $(t)(x)(y)(z)(u) (\exists v)[\beta(ztv) \Rightarrow \lambda(ytv) \ \& \ \{\beta(zvx) \vee \beta(uvx)\}]$

- P8 $(t)(x)(y)(z)(u)(\exists v)(\exists w)[\beta(xut) \ \& \ \beta(yuz) \ \& \ (x \neq u) \Rightarrow$
 $\beta(xzv) \ \& \ \beta(xyw) \ \& \ \beta(vtw)]$
- P9 $(x)(y)(z)(u)(x')(y')(z')(u')[\delta(xy'x'y') \ \& \ \delta(yzy'z') \ \& \ \delta(xux'u') \ \& \ \delta(yuy'u') \ \& \ \beta(xyz) \ \& \ \beta(x'y'z') \ \& \ (x \neq y) \Rightarrow \delta(zuz'u')]$
- P10 $(x)(y)(u)(v)(\exists z)[\beta(xyz) \ \& \ \delta(yzuv)]$
- P11 $(\exists x)(\exists y)(\exists z)[\sim \lambda(xyz)]$

It should be noted that in the presence of the other axioms, P8 is equivalent to the following axiom:

$$P8' \quad (x)(y)(z)(u)(v)[\pi(xyzu) \ \& \ \pi(xyzv) \Rightarrow \lambda(zuv)]$$

The existence axioms in \mathcal{P} guarantee the existence of those points which are the intersections of lines and those that can be constructed by the use of a "transferer of segments" (P10). If we wish to have all points which can be obtained by the use of compasses, we must add the following axiom:

$$P12 \quad (x)(y)(z)(x')(z')(u)(\exists y')[\delta(uxux') \ \& \ \delta(uzuz') \ \& \ \beta(uxy) \ \& \ \beta(xyz) \Rightarrow \delta(uyuy') \ \& \ \beta(x'y'z')]$$

This is precisely Tarski's axiom A13', and the geometry having P1–12 as axioms will be referred to as \mathcal{P}^* . It is equivalent to Tarski's system \mathcal{E}_2'' . In the presence of the remaining axioms the axiom P12 is equivalent to the axiom P12' which is stated entirely in terms of the notion β and its derived notions λ and π :

$$P12' \quad (x)(y)(z)(\exists u)(\exists v)(\exists w)[\beta(xyz) \ \& \ (x \neq y) \ \& \ (y \neq z) \Rightarrow [\lambda(xyw) \ \& \ \lambda(xuv) \ \& \ \pi(yvw) \ \& \ \pi(uvw)]]$$

If \mathfrak{F} is an ordered field, we define the (two-dimensional) coordinate geometry $\mathfrak{E}(\mathfrak{F})$ as the set of all ordered pairs $x = (x_1, x_2)$ of elements of \mathfrak{F} with the notions β and δ defined as follows:

$$\begin{aligned} \beta(xyz) &=_{\text{df}} [(x_1 - y_1)(y_2 - z_2) = (x_2 - y_2)(y_1 - z_1) \ \& \\ &\quad 0 \leq (x_1 - y_1)(y_1 - z_1) \ \& \ 0 \leq (x_2 - y_2)(y_1 - z_2)] \\ \delta(xyzu) &=_{\text{df}} [(x_1 - y_1)^2 + (x_2 - y_2)^2 = (z_1 - u_1)^2 + (z_2 - u_2)^2] \end{aligned}$$

If \mathfrak{F} has the property that the sum of two squares is a square, we call \mathfrak{F} a *Pythagorean* field. If \mathfrak{F} is a Pythagorean field, then $\mathfrak{E}(\mathfrak{F})$ is a model for \mathcal{P} .

Conversely, any model for \mathcal{P} is isomorphic to $\mathfrak{E}(\mathfrak{F})$ for some Pythagorean field \mathfrak{F} . The models for \mathcal{P}^* are isomorphic to the geometries $\mathfrak{E}(\mathfrak{F})$ where \mathfrak{F} is Euclidean, i.e. has the property that every positive element is a square. Conversely, each such geometry is a model for \mathcal{P}^* .

Elliptic geometry. One can give a similar set of axioms for elliptic plane geometry except that order is now expressed by means of a four-place relation $\gamma(xyzw)$ with the meaning that x, y, z , and w are collinear and the pair (x, y) does not separate the pair (z, w) . Again we get two systems, \mathcal{E} and \mathcal{E}^* , depending on whether or not we include the axiom corresponding to P12. This axiom is the following:

$$\text{E12 } (x)(y)(z)(w)\{\gamma(xyzw) \Leftrightarrow (\exists r)(\exists s)(\exists t)(\exists u)(\exists v)[\lambda(xyz) \ \& \ \lambda(yzw) \ \& \ \lambda(xyt) \ \& \ \lambda(xuv) \ \& \ \lambda(wrs) \ \& \ \lambda(uyr) \ \& \ \lambda(uts) \ \& \ \lambda(vtr) \ \& \ \lambda(vzs)]\}.$$

Let \mathfrak{F} be a Pythagorean field. Then by the elliptic geometry $\mathfrak{E}(\mathfrak{F})$ we mean the set of ordered triples $x = (x_1, x_2, x_3) \neq (0, 0, 0)$ from \mathfrak{F} , where (ax_1, ax_2, ax_3) is taken to be equivalent to (x_1, x_2, x_3) for $a \neq 0$. We define $\lambda(x, y, z)$ to mean the triple x, y , and z are linearly dependent; $\delta(xyzw)$ to mean that

$$\frac{(x_1y_1 + x_2y_2 + x_3y_3)^2}{(x_1^2 + x_2^2 + x_3^2)(y_1^2 + y_2^2 + y_3^2)} = \frac{(z_1w_1 + z_2w_2 + z_3w_3)^2}{(z_1^2 + z_2^2 + z_3^2)(w_1^2 + w_2^2 + w_3^2)}.$$

The notion γ can then be defined in terms of the order in \mathfrak{F} so that $\mathfrak{E}(\mathfrak{F})$ becomes a model for \mathcal{E} and all models of \mathcal{E} are isomorphic to $\mathfrak{E}(\mathfrak{F})$ for some \mathfrak{F} . The geometry $\mathfrak{E}(\mathfrak{F})$ is a model for \mathcal{E}^* if and only if \mathfrak{F} is Euclidean.

In elliptic geometry we can introduce the binary relation $\alpha(xy)$ of *polarity* between points which indicates that one point lies on the polar of the other. We can define collinearity in terms of α by the following equivalence:

$$\lambda(xyz) \Leftrightarrow (\exists t)[\alpha(tx) \ \& \ \alpha(ty) \ \& \ \alpha(tz)].$$

Hyperbolic geometry. By the elementary hyperbolic geometry \mathcal{H} we mean the geometry which follows from axioms P1–7 and P9–11 together with the negation of P8. If we assume also P12, then we call the geometry \mathcal{H}^* .

It should be remarked that the notion π which we defined for \mathcal{P} and \mathcal{P}^* here means non-intersection rather than parallelism. *Parallelism* will

be denoted by π' and is defined as follows:

$$\pi'(xyzw) =_{\text{df}} (u)(\exists v)\{\pi(xyzw) \ \& \ [\beta(xuw) \Rightarrow \beta(zuv) \ \& \ \lambda(xyv)]\}.$$

In \mathcal{H} the axiom P12 is equivalent to the following axiom which asserts the existence of parallels:

$$\text{H12 } (x)(y)(z)(\exists w)[\sim \lambda(xyz) \Rightarrow \pi'(xyzw)]$$

Let \mathfrak{F} be a Pythagorean field and e be a positive element in \mathfrak{F} such that for every $x, y \in \mathfrak{F}$ with $x^2 + y^2 < e$ there is a $z \in \mathfrak{F}$ such that $z^2 = e - x^2 - y^2$. Then a model $\mathfrak{H}(\mathfrak{F}, e)$ for \mathcal{H} is obtained by taking all pairs $x = (x_1, x_2)$ of elements from \mathfrak{F} subject to the restriction $x_1^2 + x_2^2 < e$, where the basic relations are defined by the following conditions:

$$\beta(xyz) =_{\text{df}} [(x_1 - y_1)(y_2 - z_2) = (x_2 - y_2)(y_1 - z_1) \ \&]$$

$$0 \leq (x_1 - y_1)(y_1 - z_1) \ \& \ 0 \leq (x_2 - y_2)(y_2 - z_2)],$$

and

$$\delta(xyzu) =_{\text{df}} \left[\frac{(e - x_1y_1 - x_2y_2)^2}{(e - x_1^2 - x_2^2)(e - y_1^2 - y_2^2)} = \frac{(e - z_1u_1 - z_2u_2)^2}{(e - z_1^2 - z_2^2)(e - u_1^2 - u_2^2)} \right].$$

Every model of \mathcal{H} is isomorphic to some $\mathfrak{H}(\mathfrak{F}, e)$. If e is a square then \mathfrak{F} is Euclidean and by a change of coordinates we may take $e = 1$. Every model of \mathcal{H}^* is isomorphic to $\mathfrak{H}(\mathfrak{F}) = \mathfrak{H}(\mathfrak{F}, 1)$ for some Euclidean field \mathfrak{F} .

2. Relations between order and collinearity. We have defined collinearity in our geometries in terms of order, i.e. in terms of β in the Euclidean and hyperbolic geometries and in terms of γ in the elliptic geometries. In this section we consider the possibilities of the definitions in the converse direction. The following propositions show what can be accomplished in this direction.

PROPOSITION 1. *In \mathcal{P}^* we have the following equivalence:*

$$\beta(xyz) \vee \beta(xzy) \Leftrightarrow (\exists u)(\exists v)(\exists w)[(x = y) \vee (x = z) \vee (y = z) \vee \{\lambda(xyz) \ \& \ \lambda(xyw) \ \& \ \lambda(xuv) \ \& \ \pi(uvwx) \ \& \ \pi(uvwx)\}].$$

PROPOSITION 2. *In \mathcal{E}^* we have the following equivalence¹:*

$$\gamma(xyzw) \Leftrightarrow (\exists r)(\exists s)(\exists t)(\exists u)(\exists v)[\lambda(xyz) \ \& \ \lambda(yzw) \ \& \ \lambda(xyt) \ \& \ \lambda(xuv) \ \& \ \lambda(wrs) \ \& \ \lambda(uyr) \ \& \ \lambda(uts) \ \& \ \lambda(vtr) \ \& \ \lambda(vzs)].$$

¹ This equivalence was first pointed out and used by Pieri [6] to define order in Projective Geometry!

PROPOSITION 3. *In \mathcal{H} we have the following equivalence ²:*

$$\beta(xyz) \Leftrightarrow (u)(v)(\exists w)[\lambda(xyz) \ \& \ \lambda(wyv) \ \& \ \{\lambda(wux) \vee \lambda(wuz)\}].$$

THEOREM 1. *Order can be defined in terms of collinearity in \mathcal{E}^* , \mathcal{P}^* and \mathcal{H} . On the other hand, order cannot be defined in \mathcal{E} and \mathcal{P} on the basis of collinearity and congruence.*

The possibility of defining order in \mathcal{E}^* , \mathcal{P}^* and \mathcal{H} follows from Propositions 1–3. To show the impossibility of defining order in \mathcal{E} and \mathcal{P} solely in terms of collinearity, we shall use the method of Padoa (cf. [10] and [11]) and construct the following model: Let \mathfrak{F} be the smallest field containing all algebraic numbers, an indeterminant ω , and closed under the operation of taking the square root of a sum of squares. Thus each element of \mathfrak{F} is an algebraic function $F(\omega)$ with algebraic coefficients. We make \mathfrak{F} into two distinct ordered fields \mathfrak{F}_1 and \mathfrak{F}_2 by taking two different real transcendental numbers ω_1 and ω_2 and in \mathfrak{F}_1 setting $F(\omega) > 0$ if $F(\omega_1) > 0$ and in \mathfrak{F}_2 setting $F(\omega) > 0$ if $F(\omega_2) > 0$. If we form the coordinate geometries $\mathfrak{C}(\mathfrak{F}_1)$ and $\mathfrak{C}(\mathfrak{F}_2)$ (or equivalently $\mathfrak{C}(\mathfrak{F}_1)$ and $\mathfrak{C}(\mathfrak{F}_2)$), then the natural isomorphism is a (1-1) mapping which preserves collinearity and congruence but not order.

3. The notion of orthogonality. Scott [9] has introduced the notion $\tau(xyz)$ whose meaning is that x, y , and z form a triangle with a right angle at x . This notion can be defined in terms of congruence as follows:

$$\begin{aligned} \tau(xyz) =_{\text{df}} (\exists u)(\exists v)[(u \neq y) \ \& \ (u \neq z) \ \& \ (v \neq y) \ \& \ (v \neq z) \ \& \ \delta(xy xv) \ \& \\ & \delta(xz xu) \ \& \ \delta(yz uv) \ \& \ \delta(yz zv) \ \& \ \delta(yz yu)] \end{aligned}$$

In this section we shall show that collinearity and congruence can be defined in terms of τ . For collinearity we have the following proposition:

PROPOSITION 4. *In \mathcal{E} , \mathcal{H} , and \mathcal{P} we have*

$$\lambda(xyz) \Leftrightarrow (\exists r)[\tau(rxy) \ \& \ \tau(rxz)].$$

In order to define the congruence relation δ , we introduce the auxiliary relation μ defined as follows:

$$\mu(xyz) =_{\text{df}} [\lambda(xyz) \ \& \ \delta(xy xz)].$$

² This definition of order was given by Jenks [2].

It is easy to define δ in terms of the notion of *two points being equidistant from a third* (cf. Pieri [7]). But this latter notion can be defined in terms of μ and τ by the following proposition due to Scott:

PROPOSITION 5. *In \mathcal{E} , \mathcal{P} , and \mathcal{H} we have*

$$\delta(xyyz) \Leftrightarrow (\exists r)[\mu(yxz) \vee \{\mu(rxz) \& \tau(rxy)\}].$$

Thus we can define δ from τ if we can define μ from τ . This is accomplished by the following two propositions, the first of which is due to Scott. Considerations similar to the second are found in Robinson [8], Section 3.

PROPOSITION 6. *In \mathcal{P} we have*

$$\mu(xyz) \Leftrightarrow \{[x = y \& x = z] \vee (\exists u)(\exists v)[\tau(uyz) \& \tau(vyz) \& \tau(yuv) \& \tau(zuv) \& \tau(xyu) \& \tau(xyv) \& \tau(xzu) \& \tau(xzv)]\}.$$

PROPOSITION 7. *In \mathcal{E} and \mathcal{H} we have*

$$u(xyz) \Leftrightarrow \{[x = y \& x = z] \vee (r)(\exists u)(\exists v)(\exists s)[\tau(xry) \Rightarrow \tau(xrz) \& \tau(yxu) \& \tau(zxv) \& \tau(rxu) \& \tau(rxv) \& \tau(vzs) \& \tau(uyv) \& \lambda(xrs)]\}.$$

These propositions together with the example at the end of the previous section give us the following theorem:

THEOREM 2. *In \mathcal{E} , \mathcal{H} , and \mathcal{P} the notions of collinearity λ and congruence δ can be defined in terms of τ . Thus in \mathcal{E}^* , \mathcal{P}^* , and \mathcal{H} we may use the relation τ as the sole primitive notion. On the other hand, τ does not suffice for the definition of order in \mathcal{E} and \mathcal{P} .*

4. Collinearity as the sole primitive in \mathcal{H}^* . In this section we shall show that in \mathcal{H}^* the notions of congruence can be defined in terms of collinearity λ (cf. [1], [4], and [5]). Since we have shown in the previous section that the notion τ of orthogonality can be used as the sole primitive in \mathcal{H} , it will suffice to show that τ can be defined in terms of λ .

We begin by using an auxiliary relation ψ defined as follows:

$$\psi(xyu v) =_{\text{df}} (\exists w)(\exists u')(\exists v')[\pi'(xyuw) \& \pi'(xyv'v) \& \pi'(yxuu') \& \pi'(yxvv) \& \pi'(xwu'u) \& \pi'(xwv'v)].$$

The meaning of ψ can best be seen by using a model $\mathfrak{H}(\mathfrak{F}, 1)$ of \mathcal{H}^* and

assuming without loss of generality that x has coordinates $(0, 0)$. The field \mathfrak{F} is of course a Euclidean field. The definition of $\psi(xyuv)$ then states that uv and xy are diagonals of a quadrilateral in $\mathfrak{C}(\mathfrak{F})$ whose third diagonal passes through x . Since, in the Euclidean geometry of $\mathfrak{C}(\mathfrak{F})$, x is the midpoint of the diagonal xy , we must have uv parallel to xy in the *Euclidean sense*.

With the above explanation in mind we see that for points x, y, z of $\mathfrak{S}(\mathfrak{F}, 1)$ with $x = (0, 0)$, zx will be perpendicular to xy in the Euclidean geometry of $\mathfrak{C}(\mathfrak{F})$ if and only if there are points u, v, r, s , and t such that the formulas $\psi(xyuv)$, $\psi(xzrs)$, $\pi'(uvsr)$, $\pi'(xtrs)$, and $\pi'(txvu)$ hold in $\mathfrak{S}(\mathfrak{F}, 1)$. However, the Euclidean and hyperbolic notions of a right angle coincide at the origin of $\mathfrak{C}(\mathfrak{F})$. Thus we have the following proposition:

PROPOSITION 8. *In \mathcal{H}^* we have*

$$\tau(xyz) \Leftrightarrow (\exists u)(\exists v)(\exists r)(\exists s)(\exists t)[\psi(xyuv) \& \psi(xzrs) \& \pi'(uvsr) \& \pi'(xtrs) \& \pi'(txvu)].$$

Since π' was defined in terms of λ alone, and since we have seen that τ can be used as the sole primitive notion in \mathcal{H} , we have the following corollary:

COROLLARY 3. *In \mathcal{H}^* collinearity can be taken as the sole primitive notion.*

5. Units of length. In the elliptic and hyperbolic geometries we have natural units of length, and the question immediately arises whether or not the notion of two points being at a unit distance can serve as a primitive. In the elliptic geometry the most natural distance to take is one-half the length of a straight line. We call this distance P the *polar distance*, and define the relation, $\alpha(xy)$, to mean that x and y are at distance P . This notion is easily defined in terms of congruence and collinearity, and conversely we can define orthogonality in terms of it as follows:

$$\tau(xyz) \Leftrightarrow (\exists u)(\exists v)\{\alpha(ux) \& \alpha(uz) \& \alpha(uv) \& \alpha(vx) \& \alpha(vy)\}.$$

This together with the example in Section 2 gives us the following proposition:

³ I suspect that this corollary is still true if \mathcal{H}^* is replaced by \mathcal{H} , but I have not carried out a proof. The fact that there are no parallels in a model of \mathcal{H} which is not a model for \mathcal{H}^* complicates considerations of this sort, but the method of Menger in [5] may be applicable.

PROPOSITION 9. *The binary relation α of two points being at the polar distance can be used as the sole primitive notion in \mathcal{E}^* but not in \mathcal{E} .*

On the other hand, if we use the notion $\alpha'(xy)$ of two points being at a distance less than $P/2$, we may define $\alpha(xy)$ as $\sim(\exists u)[\alpha'(xu) \& \alpha'(yu)]$. Thus in \mathcal{E} we can define collinearity and congruence in terms of α' and it is not too difficult to define order in terms of α' . Thus we have the following:

PROPOSITION 10. *In \mathcal{E} the binary notion α' of two points being closer than half the polar distance may be used as the sole primitive.*

Robinson [8] has shown that in \mathcal{P}^* with a unit of length introduced as a new primitive we cannot use the unit of length to define collinearity or congruence in elementary terms. If the points x , y , and z are within a fixed integer multiple of the unit distance of one another, then as Seidenberg has pointed out the collinearity of xyz can be defined in \mathcal{P}^* in terms of the relation of two points being at a unit distance. This definition follows from the principle of the Peaucellier invensor (cf. Robinson [8], Section 6). If we enlarge our logical basis to include finite sets of elements and add to \mathcal{P}^* the axiom of Archimedes, then we may use the unit of length as a sole primitive. Similar results hold for hyperbolic geometry.

6. Geometries with points and lines as basic elements. It follows from Robinson's results in [8] that it is not possible to find a binary relation which will serve in elementary terms as the only primitive notion for \mathcal{H}^* and \mathcal{P}^* even if we adjoin a unit of length to \mathcal{P}^* . We may use a single binary primitive notion for these geometries, however, if we cease to regard them as elementary statements about relations between points and instead regard a geometry as a class of statements about points *and* lines and relations between them. Thus, we can define a hyperbolic geometry \mathcal{H}^* which uses the single primitive ε of *incidence* between point and line. In terms of ε we define the unary notion $\rho(x)$ of *being a point* as follows:

$$\rho(x) =_{\text{df}} (y)(z)(\exists u)[\{\varepsilon(xu) \& \varepsilon(yu)\} \vee \{\varepsilon(zy) \Rightarrow \varepsilon(zu) \& \varepsilon(xu)\}].$$

With this we define collinearity among points by

$$\lambda(xyz) =_{\text{df}} (\exists w)\{\rho(x) \& \rho(y) \& \varepsilon(xw) \& \varepsilon(yw) \& \varepsilon(zw)\}.$$

We now add the axioms and definitions of \mathcal{H}^* (together with some

additional axioms to ensure that elements which are not points are lines). We then have a geometry which is isomorphic to \mathcal{H}^* when relativised to statements which only contain points as variables.

A similar procedure is possible for Euclidean geometry with a unit of length. Let $\zeta(xy)$ be the binary relation which states that *x is a point at unit distance from the line y or else y is a point at unit distance from the line x*. As before we define a point by the condition:

$$\rho(x) =_{\text{df}} (y)(z)(\exists w)[\{\zeta(xw) \ \& \ \zeta(yw)\} \vee \{\zeta(zy) \Rightarrow \zeta(zw) \ \& \ \zeta(xw)\}].$$

We can define collinearity by noting that five distinct points are collinear if they are all at a unit distance from each of two distinct lines. From this we can construct a point geometry corresponding to \mathcal{P}^* with a unit of length.

The method of Lindenbaum and Tarski [11] enables one to show that in Euclidean geometry without a unit of length there is no binary relation between points and lines from which we can define congruence.

Bibliography

- [1] ABBOTT, J. C., *The projective theory of non-Euclidean geometry*. Reports of a Mathematical Colloquium, University of Notre Dame Press (1941–1944), pp. 13–51.
- [2] JENKS, F. P., *A set of postulates for Bolyai-Lobatchevsky geometry*. Proceedings of the National Academy of Sciences, vol. 26 (1940), pp. 277–279.
- [3] MENGER, K., *Non-Euclidean geometry of joining and intersecting*. Bulletin of the American Mathematical Society, vol. 44 (1938), pp. 821–824.
- [4] ———, *A new foundation of non-Euclidean, affine, real projective and Euclidean geometry*. Proceedings of the National Academy of Sciences, vol. 24 (1938), p. 486.
- [5] ———, *New projective definitions of the concepts of hyperbolic geometry*. Reports of a Mathematical Colloquium, University of Notre Dame Press Series 2, no. 7 (1946), pp. 20–28.
- [6] PIERI, M., *I principi della geometria di posizione composti in sistema logico deduttivo*. Memorie della Reale Accademia delle Scienze di Torino, vol. 48 (1899), pp. 1–62.
- [7] ———, *La geometria elementare istituita sulle nozioni di 'punto' e 'sfera'*. Memorie di Matematica e di Fisica della Scienze, ser. 3, vol. 15 (1908), pp. 345–450.
- [8] ROBINSON, R. M., *Binary relations as primitive notions in elementary geometry*. This volume, pp. 68–85.

- [9] SCOTT, D., *A symmetric primitive notion for Euclidean geometry*. Indagationes Mathematicae, vol. 18 (1956), pp. 457–461.
- [10] TARSKI, A., *Some methodological investigations on the definability of concepts*. Logic, Semantics, Metamathematics, Oxford 1956, art. X.
- [11] ———, *On the limitations of the means of expression of deductive theories*. Logic, Semantics, Metamathematics, Oxford 1956, art. XIII (joint article with A. Lindenbaum).
- [12] ———, *What is elementary geometry?* This volume, pp. 16–29.

DIRECT INTRODUCTION OF WEIERSTRASS HOMOGENEOUS COORDINATES IN THE HYPERBOLIC PLANE, ON THE BASIS OF THE ENDCALCULUS OF HILBERT ¹

PAUL SZÁSZ

Loránd Eötvös University of Budapest, Budapest, Hungary

Introduction. In the present paper let any system of “points” and “lines” be called *hyperbolic plane* for which, besides the groups of axioms of incidence, of order and of congruence of plane I, II, III of Hilbert [3], [4] the following two axioms are valid:

AXIOM IV₁. *Let P, Q be two different points in the plane and QY a half-line on the one side of the line PQ , then there exists always one half-line PX on the same side of PQ that does not intersect QY , while every internal half-line PZ lying in the $\angle QPX$ cuts the half-line QY (Fig. 1).*

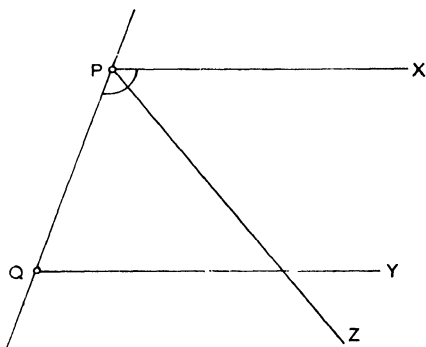


Fig. 1

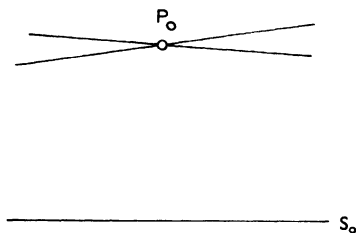


Fig. 2

AXIOM IV₂. *There exists a line s_0 and a point P_0 outside it in the plane, for which two different lines could be drawn through P_0 that do not intersect s_0 (Fig. 2).*

I have shown [7], [8] that these axioms imply the following theorem.

¹ A more detailed exposition has been published in German (see [12]).

THEOREM. *If s is an arbitrary line and P an arbitrary point outside it, then the lines drawn through P and intersecting s , form the internal lines of a certain $\angle (p_1, p_2)$ (Fig. 3). These lines p_1, p_2 , which do not intersect s any more, are called *parallels to s through P* .*

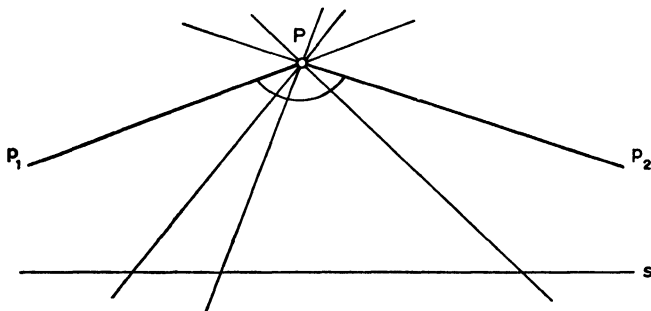


Fig. 3

This Theorem was laid down by Hilbert [3] as Axiom IV. The Axioms IV_1, IV_2 mentioned above, form together with the axiom-groups I, II, III apparently weaker assumptions than those of I, II, III, IV by the quoted author.

In the work cited above Hilbert called “ends” the points at infinity of the plane defined by any pencil of parallel lines. A line possesses, in consequence of the above Theorem, always two ends. After the proof of the fundamental theorem, according to which two lines neither intersecting each other nor being parallel, must have a common perpendicular, Hilbert was able to prove also the existence of that line which possesses two prescribed ends. From this it follows, that a determined perpendicular can be dropped on a line from an end not belonging to it. From among the preliminary theorems, stated by Hilbert for his so called *endcalculus*, I wish to stress only the one just mentioned. This endcalculus I am going to explain below, in § 1.

The way sketched by Hilbert [3] for the construction of hyperbolic geometry in the plane, leads through projective geometry. In contrast to that way there will be created in the present paper a completely *elementary* construction of hyperbolic plane geometry by means of direct introduction of certain homogeneous coordinates and an independent foundation of hyperbolic analytic geometry. Henceforth these coordinates will be called the *Weierstrass homogeneous coordinates*, because they are

identical with the well-known ones, if one assumes the axioms of continuity, instead of Axiom IV_1 , making the incomplete axiom-system complete [9]. This construction of hyperbolic geometry does not depend on hyperbolic trigonometry, the latter being a consequence of the analytic geometry of the hyperbolic plane, founded here.² Neither do I make use of Euclidean geometry, and therefore my exposition may be called an *independent* elementary foundation of hyperbolic plane geometry.

1. The endcalculus of Hilbert. The distance-function $\mathfrak{G}(t)$ and those developed from it. The endcalculus of Hilbert, somewhat altered for my purpose, follows.

Let a right angle in the plane be given with the vertex \mathcal{O} , the sides of which as half-lines have the ends Ω , E (Fig. 4).

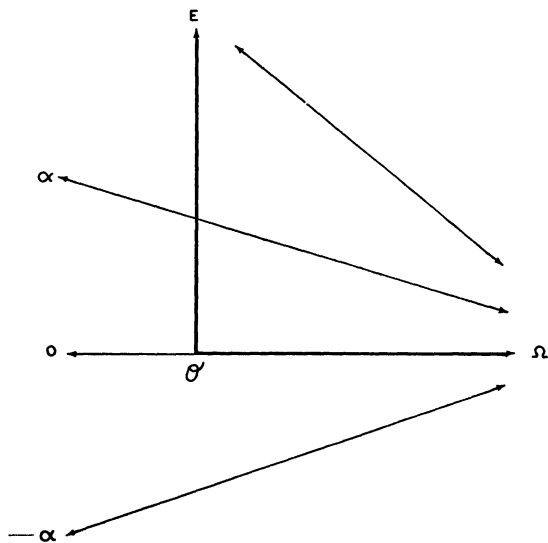


Fig. 4

The end Ω (called by Hilbert ∞) will be distinguished, and the endcalculus defined for the ends different from Ω . Such an end α should be called *positive* when the lines $\alpha\Omega$ and $E\Omega$ are lying on the same side of the line $\mathcal{O}\Omega$, and in case these lines lie on different sides of $\mathcal{O}\Omega$, the end α

² For the case of the assumption of the axioms of continuity, see Szász [10].

is called *negative*. The other end of the reflection of the line $\alpha\Omega$ in $\mathcal{O}\Omega$ should be denoted with $-\alpha$, and the other one of $\mathcal{O}\Omega$ with 0. The addition of the ends is defined by Hilbert as follows.

Let α and β be ends differing from Ω . The reflections of \mathcal{O} in $\alpha\Omega$ and $\beta\Omega$ should be denoted with \mathcal{O}_α , \mathcal{O}_β respectively (Fig. 5). The middle point of

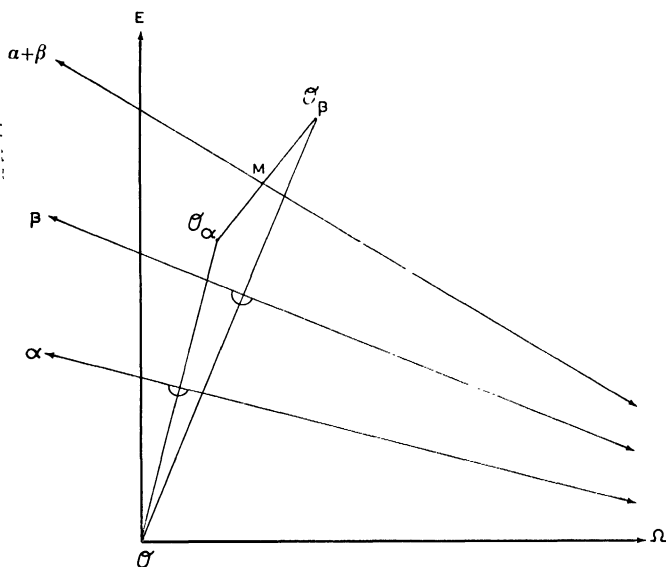


Fig. 5

the segment $\overline{\mathcal{O}_\alpha\mathcal{O}_\beta}$ being denoted with M , we define as the “sum $\alpha + \beta$ ” the other end of the line $M\Omega$.

The definition of the product might be expressed simpler by introducing, unlike Hilbert, the following distance-function that is going to be essential all through our treatment (cf. Szász [11]).

Directing the line $\mathcal{O}\Omega$ towards Ω , let us draw a perpendicular to $\mathcal{O}\Omega$ through the end-point A of the segment $\overline{\mathcal{O}A} = t$ regard being paid to sign. Let the positive end σ of this perpendicular be designated with $\mathfrak{E}(t)$:

$$(1) \quad \sigma = \mathfrak{E}(t)$$

(Fig. 6). Evidently any positive end σ corresponds to one and only one distance t with sign.

Using the designation (1) we define as the "product $\sigma_1\sigma_2$ " of the positive ends $\sigma_1 = \mathfrak{E}(t_1)$ and $\sigma_2 = \mathfrak{E}(t_2)$ the end $\mathfrak{E}(t_1 + t_2)$, i.e.

$$(2) \quad \mathfrak{E}(t_1)\mathfrak{E}(t_2) = \mathfrak{E}(t_1 + t_2)$$

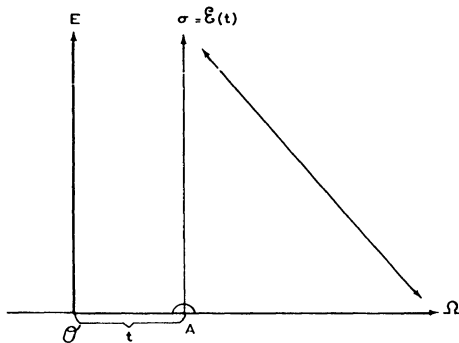


Fig. 6

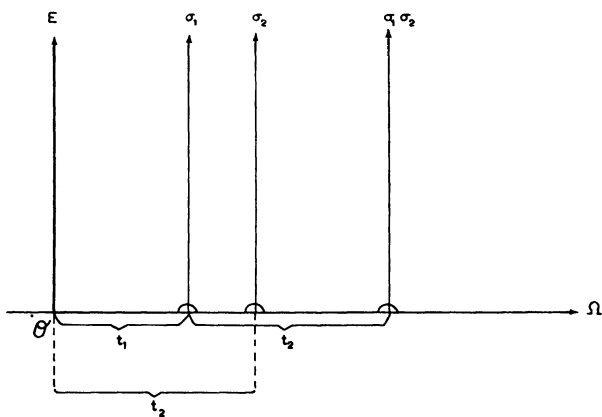


Fig. 7

(Fig. 7). Further we agree, that for positive ends α, β

$$(3) \quad \alpha(-\beta) = (-\alpha)\beta = -\alpha\beta, \quad (-\alpha)(-\beta) = \alpha\beta$$

and for any end differing from Ω , there should hold

$$(4) \quad \xi \cdot 0 = 0 \cdot \xi = 0.$$

Thus we have given the definition of the multiplication of ends differing from Ω in every case, this being equivalent to Hilbert's definition.

The positive end E is by designation (1) $\mathfrak{E}(0)$, playing the part of the *positive unity* since according to (2) $\mathfrak{E}(t)\mathfrak{E}(0) = \mathfrak{E}(0)\mathfrak{E}(t) = \mathfrak{E}(t)$. That's why we introduce the designation

$$(5) \quad \mathfrak{E}(0) = 1,$$

which by (2) may be written also as

$$(5^*) \quad \mathfrak{E}(t)\mathfrak{E}(-t) = 1.$$

The end designated with 0, which, according to (4), under multiplication plays the part of *zero*, behaves under addition also like zero, because evidently for any end differing from Ω holds

$$(6) \quad \xi + 0 = 0 + \xi = \xi$$

and

$$(7) \quad \xi + (-\xi) = 0.$$

D. Hilbert showed in his work (cited above) that *in the endcalculus defined in such a way, the familiar laws are valid concerning the four rules of arithmetic*. Or, using a modern expression: *the ends differing from Ω form a commutative field*. This field moreover has the fundamental property of *any positive end being a square*. Indeed in the sense of (2), we have

$$\mathfrak{E}(t) = \mathfrak{E}\left(\frac{t}{2}\right)^2.$$

The field of ends different from Ω , can be made an *ordered field* by the following agreement: *let α be called greater than β (β less than α) in symbols $\alpha > \beta$ ($\beta < \alpha$), in case the end $\alpha - \beta$ is positive*. One is easily convinced, that for positive ends α, β in case of $\alpha > \beta$ the line $\beta\Omega$ lies between the lines $0\Omega, \alpha\Omega$, and vice versa. From this it results that $\mathfrak{E}(t) > 1$ if $t > 0$, and then it follows at once that *in general*

$$(8) \quad \mathfrak{E}(t_2) > \mathfrak{E}(t_1), \text{ if } t_2 > t_1.$$

For the sake of brevity it is also suitable to introduce besides the distance-function $\mathfrak{E}(t)$ the following ones too:

$$(9) \quad \begin{cases} C(t) = \frac{\mathfrak{E}(t) + \mathfrak{E}(-t)}{2}, & S(t) = \frac{\mathfrak{E}(t) - \mathfrak{E}(-t)}{2}, \\ T(t) = \frac{S(t)}{C(t)} = \frac{\mathfrak{E}(t) - \mathfrak{E}(-t)}{\mathfrak{E}(t) + \mathfrak{E}(-t)}. \end{cases}$$

While $\mathfrak{E}(t)$ is the analogue of the exponential function, these latter distance-functions are the analogous of the hyperbolic functions. For the first two holds e.g. the fundamental formula

$$(10) \quad C(t)^2 - S(t)^2 = 1$$

and also the formulas

$$(11) \quad C(a + b) = C(a)C(b) + S(a)S(b)$$

and

$$(12) \quad S(a + b) = S(a)C(b) + S(b)C(a)$$

are valid for them.

Also, these distance-functions in (9) remind us of the hyperbolic functions, just as the distance-function $\mathfrak{E}(t)$ reminds us of the exponential function, e.g. it satisfies the inequality (8).

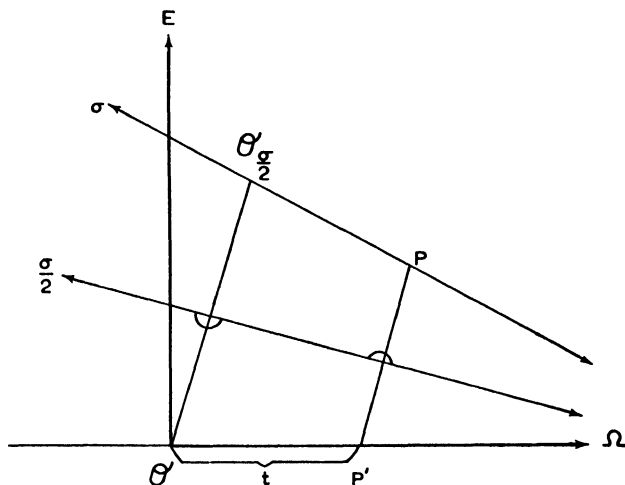


Fig. 8

2. The Weierstrass homogeneous coordinates of a point. An arbitrary point P in the plane may be characterized (Fig. 8) with the two data mentioned below. One of them is the other end of line PO , let it be σ .

However $\theta_{\frac{\sigma}{2}}$ being the reflection of θ in the line with the ends $\frac{\sigma}{2}$, Ω ,

the other end of the line $\mathcal{O}_{\frac{\sigma}{2}}\Omega$ is according to the definition of the sum of ends (§ 1), $\frac{\sigma}{2} + \frac{\sigma}{2} = \sigma$, that is to say the line $\sigma\Omega$ is the reflection of $\mathcal{O}\Omega$ in the previous line with the ends $\frac{\sigma}{2}$ and Ω . Consequently the reflection P' of P in this line joining the ends $\frac{\sigma}{2}$ and Ω , lies in $\mathcal{O}\Omega$. Now the distance $\overline{\mathcal{O}P'} = t$ taken with sign on the line $\mathcal{O}\Omega$ directed towards Ω , is the other datum, evidently determining P together with the end σ mentioned before. These data t, σ should be called *mixed-coordinates* of point P . By means of these may be proved the following

THEOREM. *The points of the hyperbolic plane and the end-triads (x_1, x_2, x_3) , built with ends differing from Ω for which holds*

$$(1) \quad x_3^2 - x_2^2 - x_1^2 = 1$$

and

$$(2) \quad x_3 > 0,$$

are put in one-to-one correspondence. This correspondence might be produced by making each point (t, σ) given in mixed-coordinates, correspond to the end-triad

$$(3) \quad \begin{cases} x_1 = S(t) + \frac{1}{2}\sigma^2\mathfrak{E}(-t) \\ x_2 = \sigma\mathfrak{E}(-t) \\ x_3 = C(t) + \frac{1}{2}\sigma^2\mathfrak{E}(-t) \end{cases}$$

The concept of inequality (§ 1) is made use of in the proof.

The ends x_1, x_2, x_3 in (3) should be called *Weierstrass homogeneous coordinates* of the point the mixed-coordinates of which are t, σ . From (3) follows for the case $t = 0, \sigma = 0$, that the *Weierstrass homogeneous coordinates of point \mathcal{O} are*

$$(4) \quad x_1 = 0, \quad x_2 = 0, \quad x_3 = 1.$$

Later on, for the transformation of the coordinates, it becomes of fundamental importance, that *for any two points (x_1, x_2, x_3) and $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ holds*

$$(5) \quad x_3\bar{x}_3 - x_2\bar{x}_2 - x_1\bar{x}_1 > 0.$$

3. The equation of the line. Weierstrass homogeneous line coordinates.

The derivation of the equation of the line may be based upon the two Lemmas of Hilbert [3] mentioned below.

LEMMA 1. α, β being ends different from Ω , the reflection of the line $\alpha\Omega$ in $\beta\Omega$ is the line joining the end $2\beta - \alpha$ with Ω .

LEMMA 2. For the ends α, β of a line that goes through \mathcal{O} holds $\alpha\beta = -1$. This plainly follows from the fact, that if from among these ends the positive one is $\alpha = \mathfrak{E}(t)$, then the other one is evidently $\beta = -\mathfrak{E}(-t)$; their product is really $-\mathfrak{E}(-t)\mathfrak{E}(t) = -\mathfrak{E}(0) = -1$.

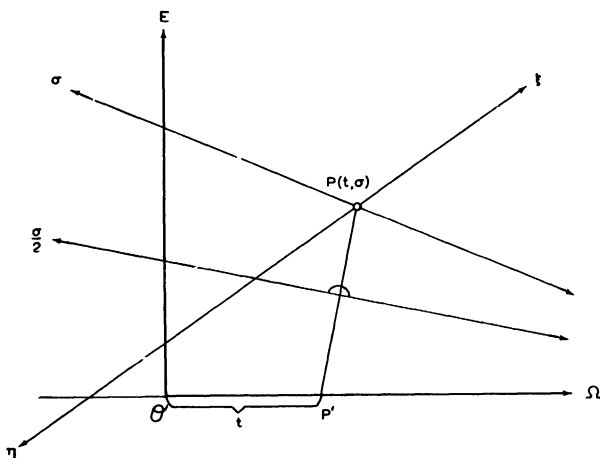


Fig. 9

Let us consider first a line possessing the ends ξ, η differing from Ω (Fig. 9). Let an arbitrary point of this line be given in mixed-coordinates (§ 2) $P(t, \sigma)$. Then by reflecting the plane in the line that joins the end $\frac{\sigma}{2}$ with Ω and after that translating it along $\mathcal{O}\Omega$ by the piece $-t$, the point P goes into \mathcal{O} . The ends ξ, η by this reflection go into the ends $\sigma - \xi, \sigma - \eta$, respectively, according to Lemma 1, and the latter ends go into $\mathfrak{E}(-t)(\sigma - \xi), \mathfrak{E}(-t)(\sigma - \eta)$, respectively, as follows from the definition of the product of ends (§ 1). Since this line goes through point \mathcal{O} already (because P is turned into \mathcal{O}), the product of these two ends due to Lemma

2 is (§ 1, (2))

$$(1) \quad \mathfrak{E}(-2t)(\sigma - \xi)(\sigma - \eta) = -1.$$

It may be seen at once, that conversely, if (1) holds for a certain point (t, σ) , then this point lies on the line $\xi\eta$. That is to say (1) is the equation of the line connecting the ends ξ, η expressed in mixed-coordinates.

Now the equation (1) can be transformed into Weierstrass homogeneous coordinates x_1, x_2, x_3 . Namely, we obtain from formulas (3) of the preceding section, by multiplying (1) with $\mathfrak{E}(t)$, that the line joining the ends ξ, η differing from Ω , has the equation

$$(2) \quad (\xi\eta - 1)x_1 + (\xi + \eta)x_2 - (\xi\eta + 1)x_3 = 0$$

in Weierstrass homogeneous coordinates.

In mixed-coordinates t, σ the equation of the line $\eta\Omega$ with the end η , is evidently $\sigma - \eta = 0$. Multiplying by $\mathfrak{E}(-t)$ and writing in terms of the coordinates x_1, x_2, x_3 we see, that the equation of the line connecting the end η with Ω is in Weierstrass homogeneous coordinates

$$(3) \quad \eta x_1 + x_2 - \eta x_3 = 0.$$

By introducing the designations

$$(4) \quad u = \frac{\xi\eta - 1}{\xi - \eta}, \quad v = \frac{\xi + \eta}{\xi - \eta}, \quad w = \frac{\xi\eta + 1}{\xi - \eta}$$

equation (2) divided by $\xi - \eta$ takes the form

$$(2^*) \quad ux_1 + vx_2 - wx_3 = 0$$

where

$$(5) \quad u^2 + v^2 - w^2 = 1.$$

The ends u, v, w in (4) should be called the *Weierstrass homogeneous line coordinates of the line $\xi\eta$ directed towards ξ* , and (2*) the *normal-form* of the equation of this line.

Per definitionem, the *Weierstrass homogeneous line coordinates of the line connecting the end η with Ω and directed towards Ω* are to be

$$(4^*) \quad u = \eta, \quad v = 1, \quad w = \eta$$

and further let (3) be the *normal-form* of the equation of this line. By reversing orientation, the line coordinates are multiplied by (-1) and the equation multiplied with (-1) should be called the normal form.

It may be easily shown, that *every equation (2*) in which (5) holds for the coefficients u, v, w is the normal-form of the equation of a certain directed line.*

4. Transformation of the Weierstrass coordinates. Let us take beside the right angle ΩOE that we have used in the definition of the endcalculus, yet another right angle $\Omega' O'E'$ where Ω' and E' are ends. Consider the congruence transformation of the plane into itself, that superposes the right angle $\Omega' O'E'$ on ΩOE . A certain directed line e should be transformed into e' by this transformation. *We mean by Weierstrass homogeneous line coordinates of the directed line e with respect to the "coordinate-system" $\Omega' O'E'$ the ones of e' with respect to the original system ΩOE .*

We define in a similar way the Weierstrass homogeneous coordinates of a point P with respect to the coordinate-system $\Omega' O'E'$.

The connection of the new coordinates with the old ones can be considered first for the line coordinates, namely by making use of the fact, that *a congruence transformation of the plane into itself, transforms every end ξ differing from Ω into the end*

$$\xi' = \frac{\alpha\xi + \beta}{\gamma\xi + \delta}$$

(it transforms in case of $\gamma \neq 0$ the end $-\frac{\delta}{\gamma}$ into Ω and this latter into the end $\frac{\alpha}{\gamma}$), where the coefficients $\alpha, \beta, \gamma, \delta$ depend only on the new system $\Omega' O'E'$ and

$$\alpha\delta - \beta\gamma = \pm 1$$

*holds, according as a correspondence in the same or in the opposite sense is involved [6].*³

On the basis of this fact, we obtain, that *the new line coordinates expressed in terms of the original ones are*

$$(1) \quad \begin{cases} u' = a_{11}u + a_{12}v + a_{13}w \\ v' = a_{21}u + a_{22}v + a_{23}w \\ w' = a_{31}u + a_{32}v + a_{33}w \end{cases}$$

where the coefficients a_{jk} depend only on the new system, and among which

³ For the proof see Gerretsen [2], Szász [11], [12]

the relations

$$(2) \quad \begin{cases} a_{11}^2 + a_{21}^2 - a_{31}^2 = 1 \\ a_{12}^2 + a_{22}^2 - a_{32}^2 = 1 \\ -a_{13}^2 - a_{23}^2 + a_{33}^2 = 1 \end{cases}$$

and

$$(3) \quad \begin{cases} a_{11}a_{12} + a_{21}a_{22} - a_{31}a_{32} = 0 \\ a_{12}a_{13} + a_{22}a_{23} - a_{32}a_{33} = 0 \\ a_{13}a_{11} + a_{23}a_{21} - a_{33}a_{31} = 0 \end{cases}$$

are valid; further, the discriminant of this transformation is

$$(4) \quad D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 1.$$

From this follows by means of a simple consideration making use of the inequality (5) of § 2, that the new coordinates of a certain point expressed in terms of the original ones are

$$(5) \quad \begin{cases} x_1' = \pm (a_{11}x_1 + a_{12}x_2 + a_{13}x_3) \\ x_2' = \pm (a_{21}x_1 + a_{22}x_2 + a_{23}x_3) \\ x_3' = \pm (a_{31}x_1 + a_{32}x_2 + a_{33}x_3) \end{cases}$$

where the coefficients a_{jk} are the same as those in (1) and the sign + or - is valid, if the two coordinate systems have the same or the opposite sense, respectively.

5. Distance of two points. The geometrical significance of the expression $u_1u_2 + v_1v_2 - w_1w_2$ for two lines. Distance of a point from a line. Choosing the new coordinate-system suitably, it follows from the formulas of the coordinate-transformation (§ 4, (5)), by means of the relations between the coefficients (§ 4, (2), (3)), that for the distance d of the points (x_1, x_2, x_3) and $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ in the original system holds

$$(1) \quad C(d) = x_3\bar{x}_3 - x_2\bar{x}_2 - x_1\bar{x}_1.$$

This formula (1) discloses the simple geometrical significance of the third coordinate x_3 at once. Namely by taking as second point \mathcal{O} the coordinates of which are 0, 0, 1 (§ 2, (4)), formula (1) expresses that the

third coordinate x_3 of a point P , determined by the distance $d = \overline{OP}$, is

$$(2) \quad x_3 = C(d).$$

Similarly, from the formulas of the transformation of the line coordinates (§ 4, (1), (2), (3)) by taking the coordinate-system suitably, follows in succession, that

(i) for the directed lines s_1, s_2 intersecting each-other, one has

$$u_1u_2 + v_1v_2 - w_1w_2 = T(a)$$

where a designates the distance with sign of the foot of the perpendicular dropped from the end of s_2 falling in the positive direction, upon s_1 (Fig. 10).

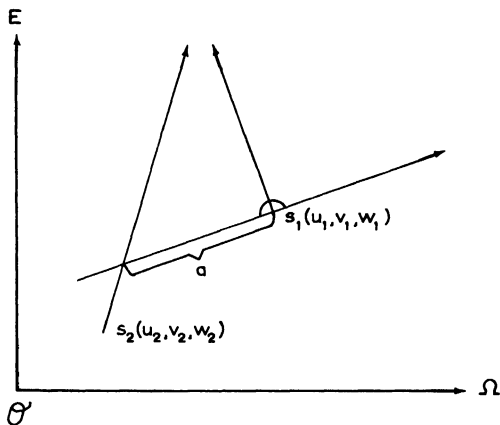


Fig. 10

(ii) for lines s_1, s_2 possessing a common perpendicular and directed equally one has

$$u_1u_2 + v_1v_2 - w_1w_2 = C(a)$$

where a signifies the piece of the common perpendicular between s_1 and s_2 (Fig. 11).

(iii) for parallel lines directed equally (Fig. 12) one has

$$u_1u_2 + v_1v_2 - w_1w_2 = 1.$$

From these theorems and the behavior of the functions $C(t)$ and $T(t)$ follows, that the lines (u_1, v_1, w_1) and (u_2, v_2, w_2) differing from each-other

1) meet if and only if

$$|u_1u_2 + v_1v_2 - w_1w_2| < 1,$$

in particular they are perpendicular if, and only if,

$$u_1u_2 + v_1v_2 - w_1w_2 = 0;$$

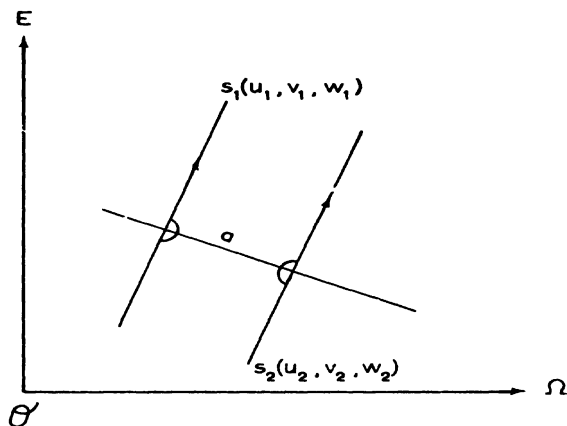


Fig. 11

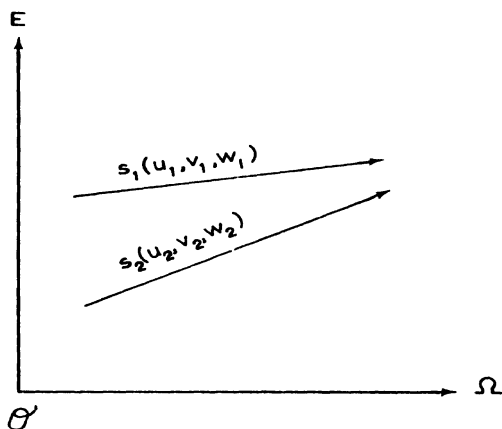


Fig. 12

2) *have a common perpendicular if, and only if,*

$$|u_1u_2 + v_1v_2 - w_1w_2| > 1;$$

3) are parallel if, and only if,

$$u_1u_2 + v_1v_2 - w_1w_2 = \pm 1.$$

Finally, it follows from a suitable choice of the new coordinate-system of the same sense, and from the formulas with respect to line and point coordinates together, that for the distance t of the point (x_1, x_2, x_3) from the directed line (u, v, w) one has

$$(3) \quad S(t) = ux_1 + vx_2 - wx_3$$

where t should be taken positive or negative, accordingly, as the point is on the positive or negative side of the line.

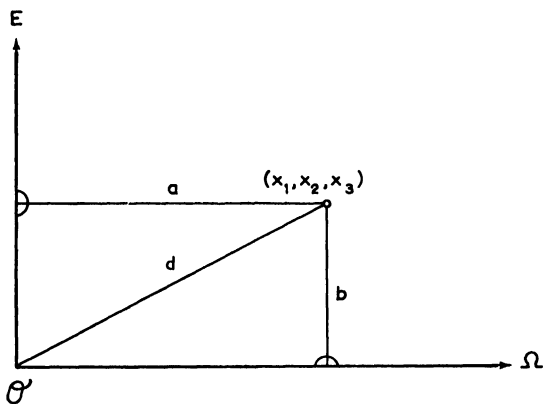


Fig. 13

This theorem discloses the simple geometrical meaning of the first two coordinates x_1, x_2 . Namely, since the end E in the endcalculus is $\xi = 1$ and the other end of the line OE is $\eta = -1$, therefore the line coordinates of this line are $-1, 0, 0$ (§ 3, (4)), thus for the signed distance $-a$ of the point (x_1, x_2, x_3) from the line OE one has by formula (3), $-x_1 = S(-a)$, or

$$(4) \quad x_1 = S(a).$$

Since moreover the line coordinates of the line $O\Omega$, directed towards Ω , are $0, 1, 0$ (§ 3, (4*)), for the signed distance b of the point (x_1, x_2, x_3) from this latter line one has by (3),

$$(5) \quad x_2 = S(b).$$

Combining these results (4) and (5) with that of (2), we may state, that *if the distance of a point from the line $\mathcal{O}E$ is a , from $\mathcal{O}\Omega$ is b , from the point \mathcal{O} is d , and we take the distance a on the right side of $\mathcal{O}E$ for positive (Fig. 13), the distance b over $\mathcal{O}\Omega$ for positive as well, and both on the other side for negative, then in the endcalculus with respect to the right angle $\Omega\mathcal{O}E$ the Weierstrass homogeneous coordinates of this point are*

$$(6) \quad x_1 = S(a), \quad x_2 = S(b), \quad x_3 = C(d).$$

The methods of § 2–5 are those by means of which I have founded, on the basis of the endcalculus of Hilbert, the analytic geometry of the hyperbolic plane. In this way I have laid the foundation for a completely elementary and at the same time independent construction of hyperbolic plane geometry.

It is not difficult indeed, on the basis of the above exposition, to introduce the homogeneous coordinates of *points at infinity* (viz. *ends*) and that of *ideal points*, further to define the concept of the *ideal line* and that of *line at infinity* analytically. The identity of hyperbolic plane geometry with the well-known circle-model of Klein-Hilbert [5], [4, p. 38] emphatically independent of continuity, is already a consequence of this analytic geometry.

To conclude we may mention that, by a result in J. C. H. Gerretsen [1], the axiom on the intersection of two circles can be derived from the axioms of the hyperbolic plane referred to at the beginning of this discussion. The analytic geometry of the hyperbolic plane outlined in the present paper provides a new proof of this result (cf. Szász [13]).

Bibliography

- [1] GERRETSEN, J. C. H., *Die Begründung der Trigonometrie in der hyperbolischen Ebene*. Koninklijke Nederlandsche Akademie van Wetenschappen, Proceedings of the Section of Sciences, vol. 45 (1942), pp. 360–366, 479–483, 559–566.
- [2] —, *Zur hyperbolischen Geometrie*. Koninklijke Nederlandsche Akademie van Wetenschappen, Proceedings of the Section of Sciences, vol. 45 (1942), pp. 567–573.
- [3] HILBERT, D., *Neue Begründung der Bolyai-Lobatschefskyschen Geometrie*. Mathematische Annalen, vol. 57 (1903), pp. 137–150.
- [4] —, *Grundlagen der Geometrie* 7. Aufl., Leipzig and Berlin 1930, pp. 159–177.

- [5] KLEIN, F., *Über die Sogenannte Nicht-Euklidische Geometrie*. Mathematische Annalen, vol. 4 (1871), pp. 583–625, spec. pp. 620–621, (reprinted in *Gesammelte Mathematische Abhandlungen I*. Berlin 1921, pp. 254–305, spec. 300–301.)
- [6] LIEBMANN, H., *Über die Begründung der hyperbolischen Geometrie*. Mathematische Annalen, vol. 59 (1904), pp. 110–128.
- [7] SZÁSZ, PAUL, *A Poincaré-féle felsik és a hiperbolikus síkgeometria kapcsolatáról* (in Hungarian). A Magyar Tudományos Akadémia III. Osztályának Közleményei, vol. 6 (1956), pp. 163–184.
- [8] —, *A remark on Hilbert's foundation of the hyperbolic plane geometry*. Acta Mathematica Academiae Scientiarum Hungaricae, vol. 9 (1958), pp. 29–31.
- [9] —, *Begründung der analytischen Geometrie der hyperbolischen Ebene mit den klassischen Hilfsmitteln, unabhängig von der Trigonometrie dieser Ebene*. Acta Mathematica Academiae Scientiarum Hungaricae, vol. 8 (1957), pp. 139–157.
- [10] —, *Die hyperbolische Trigonometrie als Folge der analytischen Geometrie der hyperbolischen Ebene*. Acta Mathematica Academiae Scientiarum Hungaricae, vol. 8 (1957), pp. 159–161.
- [11] —, *Über die Hilbertsche Begründung der hyperbolischen Geometrie*. Acta Mathematica Academiae Scientiarum Hungaricae, vol. 4 (1954), pp. 243–250.
- [12] —, *Unmittelbare Einführung Weierstrasscher homogenen Koordinaten in der hyperbolischen Ebene auf Grund der Hilbertschen Endenrechnung, Anhang*. Acta Mathematica Academiae Scientiarum Hungaricae, vol. 9 (1958), pp. 1–28, spec. 26–28.
- [13] —, *New proof of the circle axiom for two circles in the hyperbolic plane by means of the endcalculus of Hilbert*. Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös nominatae, vol. 1 (1958), pp. 97–100.

AXIOMATISCHER AUFBAU DER EBENEN ABSOLUTEN GEOMETRIE

FRIEDRICH BACHMANN

Mathematisches Seminar, Christian-Albrechts-Universität, Kiel, Deutschland

1. Absolute Geometrie soll im Sinne von J. BOLYAI als gemeinsames Fundament der euklidischen und der nichteuklidischen Geometrien verstanden werden. Die Parallelenfrage, d.h. die Frage nach dem Schneiden oder Nichtschneiden der Geraden, wird offen gelassen.

Der Aufbau der ebenen absoluten Geometrie, der hier skizziert werden soll, besitzt besonderes Interesse durch die methodische Verwendung der Spiegelungen. Anordenbarkeit und freie Beweglichkeit werden nicht gefordert. Der Begriff der absoluten Geometrie wird so allgemein gefasst, dass über allen Körpern von Charakteristik $\neq 2$, in welchen nicht jedes Element Quadrat ist, Modelle konstruiert werden können.

2. Gegeben sei zunächst eine Menge von *Punkten* und eine Menge von *Geraden*, und ferner eine *Inzidenz von Punkt und Gerade* und ein *Senkrech stehen von Geraden*, so dass die folgenden Axiome gelten:

INZIDENZAXIOME. *Es gibt wenigstens eine Gerade, und mit jeder Geraden inzidieren wenigstens drei Punkte. Zu zwei verschiedenen Punkten gibt es genau eine Gerade, welche mit beiden Punkten inzidiert.*

ORTHOGONALITÄTSAXIOME. *Ist a senkrecht zu b , so ist b senkrecht zu a . Senkrechte Geraden haben einen Punkt gemein. Durch jeden Punkt gibt es zu jeder Geraden eine Senkrechte, und wenn der Punkt mit der Geraden inzidiert, nur eine.*

SPIEGELUNGSAXIOM (SCHÜTTE). *Zu jeder Geraden g gibt es wenigstens eine Spiegelung an g , d.h. eine involutorische orthogonalitätserhaltende Kollineation, welche alle Punkte von g festlässt.*

(Eine eindeutige Abbildung einer Menge auf sich wird *involutorisch* genannt, wenn sie ihrer Umkehr-Abbildung gleich, aber von der identischen Abbildung verschieden ist).

Produkte von Geradenspiegelungen nennen wir *Bewegungen*.

Aus diesen Axiomen folgt: Den Geraden a entsprechen eindeutig die Spiegelungen σ_a an den Geraden, den Punkten A entsprechen einein-

deutig die *Punktspiegelungen* σ_A , welche dual zu den Geradenspiegelungen erklärt seien. Ferner gilt:

A, b sind *inzident* ist äquivalent mit $\sigma_A \sigma_b$ ist *involutorisch*.

a, b sind *senkrecht* ist äquivalent mit $\sigma_a \sigma_b$ ist *involutorisch*.

Indem man die Punkte und Geraden durch die Punktspiegelungen und Geradenspiegelungen, und ferner die gegebenen Relationen Inzidenz und Senkrechtstehen durch die äquivalenten Relationen zwischen den Spiegelungen ersetzt, erhält man daher in der Bewegungsgruppe ein isomorphes Abbild der gegebenen geometrischen Struktur. Dies gestattet, geometrische Sätze als Aussagen über Spiegelungen zu formulieren und durch gruppentheoretisches Rechnen mit Spiegelungen zu beweisen. Der Anwendung einer Geradenspiegelung σ_g auf die Punkte und Geraden entspricht in dem isomorphen Abbild das gruppentheoretische Transformieren aller Punkt- und Geradenspiegelungen mit der Geradenspiegelung σ_g .

3. Als *Satz von den drei Spiegelungen* bezeichnen wir die Aussage: *Das Produkt der Spiegelungen an drei Geraden a, b, c , welche mit einem Punkt inzidieren oder auf einer Geraden senkrecht stehen, ist gleich der Spiegelung an einer Geraden d .*

Eine Gesamtheit von Punkten und Geraden, für die die oben genannten Axiome und der Satz von den drei Spiegelungen gelten, werde als *metrische Ebene*, und die Theorie dieser metrischen Ebenen als *ebene absolute Geometrie* bezeichnet.

4. Für den Aufbau der ebenen absoluten Geometrie verwenden wir — entsprechend den Überlegungen in 2 — statt der bisher genannten Axiome ein Axiomensystem, welches die Bewegungsgruppen der metrischen Ebenen charakterisiert.

Wir führen zunächst einige gruppentheoretische Bezeichnungen ein. Es sei eine beliebige Gruppe gegeben. Sind α, γ Gruppenelemente, so bezeichnen wir das Element $\gamma^{-1}\alpha\gamma$, das aus α durch Transformation mit γ hervorgeht, mit α^γ . Es ist $(\alpha\beta)^\gamma = \alpha^\gamma\beta^\gamma$ und $\alpha^{\beta^\gamma} = (\alpha^\beta)^\gamma$. Eine Menge von Gruppenelementen nennen wir *invariant*, wenn sie gegen das Transformieren mit beliebigen Gruppenelementen abgeschlossen ist.

Es seien ρ, σ involutorische Gruppenelemente. Besteht für sie die Relation

$$(1) \quad \rho\sigma \text{ ist involutorisch,}$$

so schreiben wir hierfür abkürzend $\rho|\sigma$. Offenbar ist (1) äquivalent mit $\rho\sigma = \sigma\rho$ und $\rho \neq \sigma$. Wir schreiben $\rho_1, \dots, \rho_m | \sigma_1, \dots, \sigma_n$ als Abkürzung

für die Konjunktion der Aussagen $\rho_i|\sigma_k$ ($i=1, \dots, m; k=1, \dots, n$).

5 (Gruppentheoretisches Axiomensystem der ebenen absoluten Geometrie).

GRUNDANNAHME. *Es sei ein aus involutorischen Elementen bestehendes, invariantes Erzeugendensystem S einer Gruppe G gegeben.*

Die Elemente von S seien mit kleinen lateinischen Buchstaben bezeichnet. Die involutorischen Elemente aus G , welche als Produkt von zwei Elementen aus S darstellbar sind, seien mit grossen lateinischen Buchstaben (ausser G, H, S) bezeichnet.

AXIOM 1. *Zu A, B gibt es stets ein c mit $A, B|c$.*

AXIOM 2. *Aus $A, B|c, d$ folgt $A = B$ oder $c = d$.*

AXIOM 3. *Gilt $a, b, c|E$, so gibt es ein d , so dass $abc = d$ ist.*

AXIOM 4. *Gilt $a, b, c|e$, so gibt es ein d , so dass $abc = d$ ist.*

AXIOM 5. *Es gibt a, b, c derart, dass $a|b$ und weder $c|a$ noch $c|b$ noch $c|ab$ gilt.*

Dies Axiomensystem ist eine reduzierte Fassung eines von ARNOLD SCHMIDT angegebenen Axiomensystems.

6 (Gruppenebene). Ist G_m Bewegungsgruppe einer metrischen Ebene, und S_m die Menge der Geradenspiegelungen, so genügt das Paar G_m, S_m dem gruppentheoretischen Axiomensystem.

Umgekehrt lässt sich jedem Paar G, S , welches dem gruppentheoretischen Axiomensystem genügt, eine metrische Ebene durch die folgende Konstruktion der *Gruppenebene* zu G, S zuordnen:

Die Elemente a, b, \dots (die Elemente aus S) werden *Geraden*, die Elemente A, B, \dots *Punkte der Gruppenebene* genannt. Zwei Geraden a und b der Gruppenebene nennen wir zueinander *senkrecht*, wenn $a|b$ gilt. (Die Punkte sind also die Gruppenelemente, welche sich als Produkt von zwei senkrechten Geraden darstellen lassen.) Einen Punkt A und eine Gerade b der Gruppenebene nennen wir *inzident*, wenn $A|b$ gilt. Axiom 1 besagt, dass es zu zwei Punkten stets eine Verbindungsgerade gibt. Axiom 2 besagt, dass zwei verschiedene Punkte höchstens eine Verbindungsgerade besitzen. Axiom 5 spricht eine Mindest-Existenzforderung aus und besagt, dass es zwei senkrechte Geraden a, b und eine Gerade c gibt, welche weder zu a noch zu b senkrecht ist und auch nicht mit dem Punkt ab inzidiert.

Wir definieren weiter: Drei Geraden a, b, c der Gruppenebene *liegen im*

Büschel, wenn

$$(2) \quad abc \in S$$

gilt. Ist dies der Fall, gibt es also ein d mit $abc = d$, so nennen wir d die *vierte Spiegelungsgerade zu a, b, c* . Axiom 3 und Axiom 4 besagen, dass drei Geraden, welche mit einem Punkt inzidieren oder auf einer Geraden senkrecht stehen, im Büschel liegen.

Durch das Axiomensystem ist zugelassen, dass es in S Elemente a, b, c gibt, für die $abc = 1$ ist. Dann sind die Geraden a, b, c der Gruppenebene paarweise zueinander senkrecht. Wir sagen, dass drei solche Geraden ein *Polardreieck* bilden. (Polardreiecke treten bekanntlich in elliptischen Ebenen auf). Ist $abc = 1$, also $ab = c$, so ist ab als involutorisches Produkt von zwei Elementen aus S ein Element C ; es ist also dasselbe Gruppenelement sowohl Punkt als Gerade der Gruppenebene. Allgemein nennen wir, wenn $C = c$ ist, den Punkt C und die Gerade c der Gruppenebene zueinander *polar*. Ist dies der Fall, so ist jede Gerade, welche mit dem Punkt C inzidiert, zu der Geraden c senkrecht und umgekehrt; ist nämlich $C = c$, so gilt für alle x : Aus $C|x$ folgt $c|x$, und umgekehrt.

Aus den Axiomen folgt:

EXISTENZ DER SENKRECHTEN. Zu A, b gibt es stets ein c mit $A, b|c$, d.h. durch jeden Punkt gibt es zu jeder Geraden eine Senkrechte.

EINDEUTIGKEIT DER SENKRECHTEN. Aus $A, b|c, d$ folgt $A = b$ oder $c = d$, d.h. sind A, b nicht zueinander polar, so gibt es durch A nur eine Senkrechte zu b . Sind insbesondere A, b inzident, so ist das in A auf b errichtete Lot eindeutig bestimmt und gleich Ab .

Die *Spiegelung der Gruppenebene an einer Geraden c* ist die Abbildung

$$(3) \quad x^* = x^c, \quad X^* = X^c.$$

Auf Grund der Axiome 3 und 4 gilt für die Spiegelungen (3) der Satz von den drei Spiegelungen. Die *Bewegungen der Gruppenebene* sind die Abbildungen:

$$(4) \quad x^* = x^\gamma, \quad X^* = X^\gamma \quad \text{mit } \gamma \in G,$$

also die inneren Automorphismen von G , angewendet auf die Menge der Geraden und die Menge der Punkte.

Die Bewegungen (4) der Gruppenebene bilden eine Gruppe G^* , welche von der Menge S^* der Spiegelungen (3) an den Geraden der Gruppenebene erzeugt wird. Das Zentrum von G besteht nur aus dem Einselement. Das

Paar G^* , S^* ist eine Darstellung des axiomatisch gegebenen Paares G , S .

7. Sätze der absoluten Geometrie werden nun durch gruppentheoretisches Rechnen mit den involutorischen Elementen a, b, \dots und A, B, \dots bewiesen. Es gibt mancherlei einfache Beweise dieser Art.

Als Beispiel betrachten wir den Satz von der isogonalen Verwandtschaft in bezug auf ein Dreieit a, b, c . Er kann folgendermassen formuliert werden: Sind a', b', c' Geraden, welche im Büschel liegen, und liegen b, a', c sowie c, b', a sowie a, c', b im Büschel, so liegen auch die vierten Spiegelungsgeraden $ba'c = a''$, $cb'a = b''$, $ac'b = c''$ im Büschel.

SATZ VON DER ISOGONALEN VERWANDTSCHAFT. Aus $ba'c = a''$, $cb'a = b''$, $ac'b = c''$ und $a'b'c' \in S$ folgt $a''b''c'' \in S$.

BEWEIS. Es ist $a''b''c'' = ba'c \cdot cb'a \cdot ac'b = (a'b'c')^b$. Aus $a'b'c' \in S$ folgt $(a'b'c')^b \in S$, wegen der Invarianz von S , und damit $a''b''c'' \in S$.

Zu der dreistelligen Relation (2), durch die das Im-Büschel-Liegen von Geraden erklärt ist, bemerken wir:

Wegen der Invarianz von S ist die Relation (2) *reflexiv* und *symmetrisch* in dem folgenden Sinne: Für Elemente a, b, c , die nicht sämtlich verschieden sind, gilt (2) stets; gilt (2) für Elemente a, b, c , so auch für jede Permutation von a, b, c . Aus dem Axiomensystem der absoluten Geometrie folgt, dass die Relation (2) auch *transitiv* ist, d.h. der

TRANSITIVITÄTSSATZ. Aus $a \neq b$ und $abc, abd \in S$ folgt $acd \in S$.

Nützlich für das Beweisen in der absoluten Geometrie sind Lemmata über nicht notwendig involutorische Elemente aus G , wie die folgenden:

LEMMA VON THOMSEN. α und β seien Elemente aus G , welche als Produkte einer ungeraden Anzahl von Elementen aus S darstellbar sind. Ist $\alpha \neq 1$ und $\alpha^\beta = \alpha^{-1}$, so liegt α oder β in S .

	β_1	β_2	β_3	<p>LEMMA VON DEN NEUN INVOLUTORISCHEN PRODUKTEN. Sind $\alpha_i, \beta_k \in G$ ($i, k = 1, 2, 3$) und $\alpha_1 \neq \alpha_2$, $\beta_1 \neq \beta_2$, so gilt: Steht an den acht mit \circ bezeichneten Stellen der Produkttafel der $\alpha_i\beta_k$ ein Element aus S, so auch an der mit $*$ bezeichneten Stelle.</p>
α_1	\circ	\circ	\circ	
α_2	\circ	\circ	\circ	
α_3	\circ	\circ	$*$	

Aus dem Lemma von THOMSEN erhält man durch Einsetzung z.B. den Höhensatz, aus dem Lemma von den neun involutorischen Produkten z.B. den HESSENBERGSchen Gegenpaarungssatz (Vierseitsatz), mit dem sich der Satz von PAPPUS gewinnen lässt.

Als Beispiel sei etwa der Beweis des Höhensatzes hier ausgeführt. Wir

betrachten ein Dreiseit, welches kein Polardreiseit ist und dessen Seiten nicht im Büschel liegen. Unter einer Höhe verstehen wir eine Gerade, welche auf einer Seite des Dreiseits senkrecht steht und mit den beiden anderen Seiten im Büschel liegt.

HÖHENSATZ. Ist $abc \neq 1$ und $abc \notin S$ und gilt:

$$(5) \quad u|a, \quad v|b, \quad w|c,$$

$$(6) \quad bcu, \quad cav, \quad abw \in S,$$

so ist $uvw \in S$.

BEWEIS. Nach der ersten Voraussetzung (5) ist $ua = au$, also $a^u = a$, und nach der ersten Voraussetzung (6) $bcu = uc b$, also $(bc)^u = cb$, insgesamt also $(abc)^u = a^u(bc)^u = acb$. Indem man den Schluss wiederholt, erhält man

$$(abc)^{uvw} = (a^u(bc)^u)^{vw} = (acb)^{vw} = ((ac)^{vb})^w = (cab)^w = c^w(ab)^w = cba.$$

Es ist also

$$(abc)^{uvw} = (abc)^{-1},$$

und hieraus folgt wegen $abc \neq 1$ und $abc \notin S$ nach dem Lemma von THOMSEN die Behauptung.

8 (Geradenbüschel). Da die dreistellige Relation (2), wie in 7 bemerkt, reflexiv, symmetrisch und transitiv ist, definiert sie in S Teilmengen mit den Eigenschaften: 1) Für je drei Elemente a, b, c einer Teilmenge gilt (2); 2) Besteht zwischen zwei verschiedenen Elementen a, b einer Teilmenge und einem Element c die Relation (2), so gehört auch c der Teilmenge an; 3) Zu je zwei Elementen a, b gibt es eine Teilmenge, der sie angehören. Aus 1), 2), 3) folgt: Je zwei verschiedene Teilmengen haben höchstens ein Element gemein.

Diese durch die Relation (2) definierten Teilmengen der Menge aller Geraden nennen wir *Geradenbüschel*. Je zwei verschiedene Geraden a, b bestimmen ein Geradenbüschel; es besteht aus allen Geraden c , die mit a, b im Büschel liegen.

Alle Geraden, welche mit einem gegebenen Punkt A inzidieren, bilden ein Geradenbüschel, das wir mit $G(A)$ bezeichnen. Solche Geradenbüschel nennen wir *eigentliche Geradenbüschel*.

Alle Geraden, welche auf einer gegebenen Geraden a senkrecht stehen, bilden ein Geradenbüschel, das *Lotbüschel* zu a , das wir mit $G(a)$ bezeichnen.

9 (Halbdrehungen). Ein weiteres Hilfsmittel für Überlegungen in

der absoluten Geometrie sind gewisse Abbildungen, welche keine Bewegungen sind, nämlich die von HJELMSLEV eingeführten Halbdrehungen.

Jedes Element α aus G , welches als Produkt einer ungeraden Anzahl von Elementen aus S darstellbar ist, lässt sich in der Form

$$(7) \quad abc \text{ mit } a|b,c$$

darstellen. Ist $\alpha \neq 1$, so bestimmt α das in der Darstellung (7) auftretende Element a , das wir mit $[\alpha]$ bezeichnen, eindeutig.

Es sei nun γ ein nicht-involutorisches Element aus G , welches als Produkt von zwei, mit einem Punkt O inzidierenden Geraden darstellbar ist. Durch $x \rightarrow [x\gamma]$ wird eine eineindeutige Abbildung der Menge der Geraden der Gruppenebene in sich definiert. Diese Abbildung nennen wir die *Halbdrehung um O , welche zu dem Gruppenelement γ gehört*, und bezeichnen sie mit H_γ . Es ist also

$$(8) \quad xH_\gamma = [x\gamma].$$

Die Halbdrehungen sind büscheltreu: Liegen drei Geraden im Büschel, so liegen auch ihre Bildgeraden im Büschel, und umgekehrt. Insbesondere bildet jede Halbdrehung um O die Menge der Geraden durch O eineindeutig auf sich ab. Senkrechte Geraden werden im allgemeinen nicht in senkrechte Geraden übergehen, wohl aber dann, wenn eine der beiden Geraden durch O geht.

Jede Halbdrehung induziert eine eineindeutige Abbildung der Menge der Geradenbüschel auf sich; dabei wird die Menge der eigentlichen Geradenbüschel in sich abgebildet. Wir nennen auch diese Abbildung der Geradenbüschel eine Halbdrehung und werden sie mit dem gleichen Symbol bezeichnen, wie die Halbdrehung der Geraden, durch die sie induziert wird. Die Menge der Lotbüschel der Geraden durch O wird durch jede Halbdrehung um O auf sich abgebildet; jedes andere Geradenbüschel kann durch eine geeignete Halbdrehung um O in ein eigentliches Geradenbüschel übergeführt werden.

10. Aus gewissen Sätzen der absoluten Geometrie entstehen bei gewissen Ersetzungen von Punkten durch Geraden oder von Geraden durch Punkte wieder richtige Sätze der absoluten Geometrie. Ein Beispiel für diese „*Punkt-Geraden-Analogie*“, auf die ARNOLD SCHMIDT aufmerksam gemacht hat, sind Axiom 3 und Axiom 4; weitere Beispiele sind:

Zu A, B gibt es stets ein c mit $A, B|c$
(Existenz der Verbindungsgeraden),

Zu A, b gibt es stets ein c mit $A, b|c$
(Existenz der Senkrechten).

Aus $A, B c, d$ folgt $A = B$ oder $c = d$ (Eindeutigkeit der Verbindungsgeraden).	Aus $A, b c, d$ folgt $A = b$ oder $c = d$ (Eindeutigkeit der Senkrechten).
---------------------------------------------------------------------------------------	--------------------------------------------------------------------------------

Ersetzt man in den rechts stehenden Sätzen auch den Punkt A durch eine Gerade a , so erhält man die Aussagen

V Zu a, b gibt es stets ein c mit $a, b|c$,

d.h. je zwei Geraden haben ein gemeinsames Lot, und

$\sim R$ Aus $a, b|c, d$ folgt $a = b$ oder $c = d$,

d.h. zwei verschiedene Geraden haben höchstens ein gemeinsames Lot.

Die Aussage $\sim R$ ist die Negation der Aussage

R Es gibt a, b, c, d mit $a, b|c, d$ und $a \neq b$ und $c \neq d$,

welche besagt, dass ein Rechtseit existiert.

Keine von den Aussagen V, $\sim R$, R ist aus den Axiomen der absoluten Geometrie beweisbar. Man kann jede von ihnen als ein Zusatzaxiom zu dem Axiomensystem aus 5 hinzufügen und so Spezialfälle der absoluten Geometrie definieren. Die Aussage V ist mit der Existenz von Polar-dreiseiten äquivalent und definiert die *elliptische Geometrie* im Rahmen unseres Axiomensystems der absoluten Geometrie. Die Aussage R nennen wir das *Axiom der euklidischen Metrik*, die Aussage $\sim R$ das *Axiom der nichteuklidischen Metrik*. Die Zusatzaxiome R und $\sim R$ führen zu der Gabelung der absoluten Geometrie in die *Geometrie mit euklidischer Metrik* und die *Geometrie mit nichteuklidischer Metrik*. Aus V folgt $\sim R$.

Ein allgemeines Theorem, welches den Umfang der in der absoluten Geometrie erlaubten Ersetzungen von Punkten durch Geraden und von Geraden durch Punkte beschreibt, ist nicht bekannt. Jedoch sind in der durch das Zusatzaxiom V definierten elliptischen Geometrie beliebige Ersetzungen dieser Art erlaubt.

11 (Projektiv-metrische Ebenen). Unter einer *projektiven Ebene* verstehen wir eine Menge von Punkten und Geraden, in der die projektiven Inzidenzaxiome, der Satz von PAPPUS und das FANO-Axiom gelten.

Eine projektive Ebene, in der eine Gerade als „unendlichferne“ Gerade g_∞ und auf ihr eine projektive fixpunktfreie Involution als „absolute“ Involution ausgezeichnet ist, nennen wir eine *singuläre projektiv-metrische Ebene*. Jeder Geraden $a \neq g_\infty$ ordnen wir einen *Pol* zu, nämlich den auf g_∞ liegenden Punkt, welcher dem Schnittpunkt von a, g_∞ in der absoluten Involution entspricht.

Eine projektive Ebene, in der eine projektive Polarität als „absolute“ Polarität ausgezeichnet ist, nennen wir eine *ordinäre projektiv-metrische Ebene*.

Es sei nun c eine Gerade einer gegebenen projektiv-metrischen Ebene; die Gerade c sei im singulären Fall von g_∞ verschieden und im ordinären Fall nicht mit ihrem Pol inzident. Dann nennen wir die harmonische Homologie, deren Achse die Gerade c und deren Zentrum der Pol von c ist, die *Spiegelung der projektiv-metrischen Ebene an der Geraden c* . Die von der Menge S_{pm} dieser Spiegelungen an Geraden der projektiv-metrischen Ebene erzeugte Gruppe G_{pm} nennen wir die *Bewegungsgruppe der projektiv-metrischen Ebene*.

12 (Idealebene). Die Gruppenebene zu G , S lässt sich durch Einführung von idealen Elementen zu einer projektiv-metrischen Ebene erweitern.

Man nennt hierzu die Geradenbüschel *Idealpunkte*, und die eigentlichen Geradenbüschel *eigentliche Idealpunkte*. Die Menge aller Geradenbüschel, welche eine Gerade a gemein haben, bezeichnet man als die *eigentliche Idealgerade $g(a)$* .

Um den Begriff der Idealgeraden allgemein zu definieren, verwenden wir die Halbdrehungen, die es ermöglichen, „Uneigentliches“ in „Eigentliches“ überzuführen (vgl. 9).

Wir wählen einen Punkt O der Gruppenebene, den wir fortan festhalten. Eine Halbdrehung H_γ um O führt jede eigentliche Idealgerade in eine eigentliche Idealgerade über; denn es ist

$$(9) \quad g(a)H_\gamma = g(aH_\gamma).$$

Die Menge der Lotbüschel der Geraden durch O , die bei jeder Halbdrehung um O in sich übergeht, bezeichnen wir mit $g(O)$.

Eine Menge a von Idealpunkten wird nun eine *Idealgerade* genannt, 1) wenn es eine Halbdrehung H_γ um O gibt, so dass aH_γ eine eigentliche Idealgerade ist, und ferner 2) wenn $a = g(O)$ ist.

Man beweist dann, dass die Idealpunkte und Idealgeraden eine projektive Ebene bilden, die *Idealebene zu G , S* . Die eigentlichen Idealpunkte und die eigentlichen Idealgeraden bilden eine zu der Gruppenebene isomorphe Teilebene der Idealebene.

Es ist nun zu zeigen, dass die in der Gruppenebene erklärte Orthogonalität in der Idealebene projektiv-metrische Relationen induziert.

Wir nehmen zunächst an, dass in der Gruppenebene das Axiom der euklidischen Metrik gilt. Dann sind je zwei Geraden, welche ein gemein-

sames Lot haben, zueinander *lotgleich*, d.h. es ist jedes Lot der einen Geraden auch Lot der anderen Geraden. Daher ist jedes Lotbüschel auch Lotbüschel einer Geraden durch einen fest gewählten Punkt. Die Menge aller Lotbüschel ist also eine Idealgerade, die wir mit g_∞ bezeichnen.

Gibt es in einem Lotbüschel eine Gerade, welche zu einer Geraden eines anderen Lotbüschels orthogonal ist, so ist jede Gerade des einen Lotbüschels zu jeder Geraden des anderen Lotbüschels orthogonal. Es gibt daher eine *Orthogonalität der Lotbüschel*. Sie definiert auf der ausgezeichneten Idealgeraden g_∞ eine projektive fixelementfreie Involution. Die Idealebene einer Gruppenebene mit euklidischer Metrik ist also eine singuläre projektiv-metrische Ebene.

Es gelte nun in der Gruppenebene das Axiom der nichteuklidischen Metrik. Dann ist jede Gerade nur zu sich selbst lotgleich, und die Lotbüschel verschiedener Geraden sind verschieden. Ordnet man jeder eigentlichen Idealgeraden $g(a)$ den Idealpunkt $G(a)$ (das Lotbüschel von a) als Pol zu, so ist dies jetzt eine eindeutige Zuordnung zwischen den eigentlichen Idealgeraden und den Lotbüscheln. Um diese Zuordnung zu einer in der gesamten Idealebene erklärten Polarität auszudehnen, verwenden wir wiederum die Halbdrehungen um den Punkt O . Wendet man zunächst auf eine eigentliche Idealgerade $g(a)$ eine Halbdrehung H_γ um O an, so entsteht nach (9) die eigentliche Idealgerade $g(aH_\gamma)$. Der Pol $G(a)$ von $g(a)$ wird dabei im allgemeinen nicht wieder in den Pol $G(aH_\gamma)$ von $g(aH_\gamma)$ übergehen. Vielmehr besteht zwischen $G(a)$ und $G(aH_\gamma)$ der folgende allgemeine Zusammenhang: Es ist

$$(10) \quad G(a)H_{\gamma^{-1}}^{-1} = G(aH_\gamma).$$

Wir nennen jedes Paar $g(a)$, $G(a)$ ein *primitives Polare-Pol-Paar* und definieren nun für eine Idealgerade a und einen Idealpunkt A :

a , A heißen ein *Polare-Pol-Paar*, 1) wenn es eine Halbdrehung H_γ um O gibt, so dass aH_γ , $AH_{\gamma^{-1}}^{-1}$ ein primitives Polare-Pol-Paar sind, und ferner 2) wenn $a = g(O)$, $A = G(O)$ ist.

Man beweist nun, dass hiermit in der Idealebene eine projektive Polarität erklärt ist. Die Idealebene einer Gruppenebene mit nichteuklidischer Metrik ist also eine ordinäre projektiv-metrische Ebene.

13. Die Spiegelung (3) der Gruppenebene an einer Geraden c induziert die Spiegelung der projektiv-metrischen Idealebene an der eigentlichen Idealgeraden $g(c)$. Die Bewegungen (4) der Gruppenebene induzieren daher Bewegungen der projektiv-metrischen Idealebene. Damit ergibt sich nun das

HAUPTTHEOREM. *Jedes Paar G, S , welches dem gruppentheoretischen Axiomensystem aus 5 genügt, lässt sich als Teilsystem eines Paares G_{pm}, S_{pm} darstellen.*

Anders gesagt: Die Bewegungsgruppen der metrischen Ebenen sind als Untergruppen von Bewegungsgruppen projektiv-metrischer Ebenen darstellbar.

14 (Metrische Vektorräume und orthogonale Gruppen). Sei $V_3(K, F)$ der durch eine symmetrische bilineare Form F metrisierte dreidimensionale Vektorraum über einem Körper K von Charakteristik $\neq 2$. Wenn in dem metrischen Vektorraum $V_3(K, F)$ alle isotropen Vektoren im Radi- kal liegen, wird die Form F *nullteilig* genannt.

Die *eigentlich-orthogonale Gruppe* $O_3^+(K, F)$ wird erklärt als die Gruppe aller linearen Abbildungen des metrischen Vektorraumes $V_3(K, F)$ auf sich, welche den Wert von F erhalten und die Determinante 1 haben. Unter der *Spiegelung des metrischen Vektorraumes an einem nicht-isotropen eindimensionalen Teilraum T* verstehen wir die involutorische lineare Abbildung des metrischen Vektorraumes auf sich, welche jeden Vektor des Teilraumes T festlässt und jeden Vektor des orthogonalen Komplements von T in den entgegengesetzten überführt. Die Menge $S_3^+(K, F)$ aller dieser Spiegelungen des metrischen Vektorraumes ist ein Erzeugendensystem der Gruppe $O_3^+(K, F)$.

15. Jede projektive Ebene kann man als dreidimensionalen Vektorraum über einem Körper K von Charakteristik $\neq 2$ darstellen, indem man die Geraden durch die eindimensionalen und die Punkte durch die zweidimensionalen Teilräume des Vektorraumes darstellt. Jede projektiv-metrische Ebene kann man in entsprechender Weise als metrischen Vektorraum $V_3(K, F)$ darstellen; die Form F ist im singulären Fall vom Rang 2 und nullteilig, im ordinären Fall vom Rang 3. Die Spiegelungen der projektiv-metrischen Ebene an den in 11 genannten Geraden lassen sich durch die Spiegelungen des metrischen Vektorraumes an den nicht-isotropen eindimensionalen Teilräumen darstellen. Für die Bewegungsgruppe der projektiv-metrischen Ebene gilt daher: Das Paar G_{pm}, S_{pm} kann dargestellt werden durch das Paar

$$(11) \quad O_3^+(K, F), \quad S_3^+(K, F).$$

Das Haupttheorem gestattet daher, die Gruppen, welche das Axiomensystem der absoluten Geometrie erfüllen, — anders gesagt, die Bewegungsgruppen der axiomatisch gegebenen metrischen Ebenen — als

Gruppen von orthogonalen Transformationen metrischer Vektorräume darzustellen:

HAUPTTHEOREM, algebraische Fassung. *Jedes Paar G, S , welches dem Axiomensystem aus 5 genügt, ist darstellbar als Teilsystem eines Paares (11), wobei der Körper K von Charakteristik $\neq 2$ und die symmetrische bilineare Form F vom Rang 2 und nullteilig oder vom Rang 3 ist.*

16. Umgekehrt entsteht nun die Frage, welche Teilsysteme von solchen Paaren (11) Modelle des Axiomensystems aus 5 sind. Hierzu sei hier folgendes gesagt:

FALL 1: F vom Rang 2 und nullteilig (euklidische Metrik). In diesem Fall genügt jedes Paar (11) dem Axiomensystem. Gibt es in dem durch F metrisierten Vektorraum zwei orthogonale Einheitsvektoren, so lassen sich in jedem Paar (11) alle „zugehörigen“, dem Axiomensystem genügenden Teilsysteme algebraisch beschreiben. Dabei spielt der von den Elementen $(1 + c^2)^{-1}$ mit $c \in K$ erzeugte Teilring von K eine Rolle.

FALL 2: F vom Rang 3 und nullteilig (elliptische Metrik). Auch in diesem Falle genügt jedes Paar (11) dem Axiomensystem. Beispiele von echten Teilsystemen, welche dem Axiomensystem genügen, sind bekannt; eine allgemeine Charakterisierung scheint schwieriger als im Fall 1.

FALL 3: F vom Rang 3 und nicht nullteilig (hyperbolische Metrik). In diesem Fall genügt kein Paar (11) dem Axiomensystem. Jedoch kann es ein echtes Teilsystem S von $S_3^+(K, F)$ geben, so dass das Paar $O_3^+(K, F)$, S dem Axiomensystem genügt. Wird F so normiert, dass die Determinante von F diejenige Quadratklasse von K ist, der die 1 angehört, so gilt: Ist K geordnet, und S die Menge der Elemente aus $S_3^+(K, F)$ mit negativer Norm, so genügt $O_3^+(K, F)$, S dem Axiomensystem. Es sind alle invarianten, und Beispiele nicht-invarianter Teilsysteme der Paare (11) bekannt, welche dem Axiomensystem genügen.

Bibliographie

- BOLYAI, J., *Appendix. Scientiam spatii absolute veram exhibens: a veritate aut falsitate Axiomatis XI Euclidei (a priori haud unquam decidenda) independentem: adjecta ad casum falsitatis, quadratura circuli geometrica.* Maros-Vasarhely 1832.
- WIENER, H., *Die Zusammensetzung zweier endlicher Schraubungen zu einer einzigen. Zur Theorie der Umwendungen. Über geometrische Analysen. Über geometrische*

Analysen, Fortsetzung. Über die aus zwei Spiegelungen zusammengesetzten Verwandtschaften. Über Gruppen vertauschbarer zweispiegeliger Verwandtschaften. Berichte über die Verhandlungen der Kgl. Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Mathematisch-naturwissenschaftliche Klasse. Band 42 (1890), S. 13–23, 71–87, 245–267; Band 43 (1891), S. 424–447, 644–673; Band 45 (1893), S. 555–598.

DEHN, M., *Die Legendreschen Sätze über die Winkelsumme im Dreieck.* Mathematische Annalen. Band 53 (1900), S. 404–439.

HESSENBERG, G., *Neue Begründung der Sphärik.* Sitzungsberichte der Berliner Mathematischen Gesellschaft. Band 4 (1905), S. 69–77.

HJELMSLEV, J., *Neue Begründung der ebenen Geometrie.* Mathematische Annalen. Band 64 (1907), S. 449–474.

SCHUR, F., *Grundlagen der Geometrie.* Leipzig 1909. X + 192 S.

HJELMSLEV, J., *Einleitung in die allgemeine Kongruenzlehre.* Det Kgl. Danske Videnskabernes Selskab, Matematisk-fysiske Meddelelser. Band 8 (1929), Nr. 11. Band 10 (1929), Nr. 1. Band 19 (1942), Nr. 12. Band 22 (1945), Nr. 6. Band 22 (1945), Nr. 13. Band 25 (1949), Nr. 10.

HESSENBERG, G., *Grundlagen der Geometrie.* Berlin/Leipzig 1930. 143 S.

THOMSEN, G., *Grundlagen der Elementargeometrie in gruppentheoretischer Behandlung.* Hamburger Mathematische Einzelschriften. Heft 15. Leipzig/Berlin 1933. 88 S.

REIDEMEISTER, K., *Geometria proiettiva non euclidea.* Rendiconti del Seminario Mathematico della R. Università di Roma. Serie III, volume 1, parte 2 (1934), p. 219–228.

BACHMANN, F., *Eine Begründung der absoluten Geometrie in der Ebene.* Mathematische Annalen. Band 113 (1936), S. 424–451.

SCHMIDT, ARNOLD, *Die Dualität von Inzidenz und Senkrechtstehen in der absoluten Geometrie.* Mathematische Annalen. Band 118 (1943), S. 609–635.

SERNER, E., *Ein gruppentheoretischer Beweis des Satzes von Desargues in der absoluten Axiomatik.* Archiv der Mathematik. Band 5 (1954), S. 458–468.

SCHÜTTE, K., *Die Winkelmetrik in der affin-orthogonalen Ebene.* Mathematische Annalen. Band 130 (1955), S. 183–195.

Gruppentheoretisches Axiomensystem einer verallgemeinerten euklidischen Geometrie. Mathematische Annalen. Band 132 (1956), S. 43–62.

BACHMANN, F., *Aufbau der Geometrie aus dem Spiegelungsbegriff.* Die Grundlehren der mathematischen Wissenschaften. Band 96. Berlin/Göttingen/Heidelberg. 1959. XIV + 312 S.

[In dem an letzter Stelle genannten Buch ist der hier skizzierte axiomatische Aufbau der ebenen absoluten Geometrie durchgeführt.]

NEW METRIC POSTULATES FOR ELLIPTIC n -SPACE

LEONARD M. BLUMENTHAL

University of Missouri, Columbia, Missouri, U.S.A.

1. Introduction. In its most general aspects, a *distance space* is formed from an abstract set S by mapping the set of all ordered pairs of elements of S into a second set, which may be a subset of S . It is suggestive to call the elements of S *points*, and the elements of the second set *distances*. Distance spaces are particularized by specifying the distance sets and by postulating properties of the mapping. If, for example, the distance set is the class of non-negative real numbers, and the mapping that associates with each pair p, q of elements of the set S the number pq is *definite* (that is, $pq = 0$ if and only if $p = q$), and *symmetric* ($pq = qp$), the resulting distance space is called *semimetric*. The class of *metric* spaces is obtained by assuming, in addition, that if $p, q, r \in S$, the associated distances pq, qr, pr satisfy the *triangle inequality*, $pq + qr \geq pr$. For each positive integer n , the classical spaces (euclidean, spherical, hyperbolic, and elliptic) of n dimensions are metric spaces.

A given distance space \mathcal{Z} is characterized metrically with respect to a prescribed class of distance spaces when necessary and sufficient conditions, *expressed wholly and explicitly in terms of the distance*, are formulated in order that any member of the class may be mapped onto \mathcal{Z} in a distance-preserving manner. A mapping of this kind is called a congruence. It is clear that such a metric characterization induces an axiomatization of \mathcal{Z} in terms of the sole (geometric) primitive notions of point and distance when the given class of comparison spaces is sufficiently general.

Euclidean spaces R_n were the first to be studied in this manner. In his *Zweite Untersuchung*, Menger obtained metric postulates for euclidean n -space by first solving the more general problem of characterizing metrically subsets of R_n , with respect to the class of semimetric spaces [6]. With this accomplished, the solution of the space problem follows upon adjoining to the metric characterization of its subsets (with respect to the class of semimetric spaces) those metric properties that serve to distinguish the R_n itself among its subsets. It was noted by W. A. Wilson, however, that though none of Menger's conditions for congruently

imbedding an arbitrary semimetric space into the R_n can be suppressed, the set of assumptions obtained by adjoining to those conditions the properties that individualize the R_n among the subsets (needed to characterize the whole R_n) can be very materially reduced [8]. Wilson's reduction consists in replacing Menger's assumption that for every integer k , ($1 \leq k \leq n$), each $(k + 1)$ -tuple of points of a semimetric space can be congruently imbedded in R_n , by the much milder requirement that each four points be imbeddable in R_3 . The *crucial* imbedding sets are thus *quadruples* of points, regardless of the dimension of the euclidean space being characterized.

The following comments concerning Wilson's contribution are pertinent. (1). In validating the sufficiency of his "four-point" property, Wilson made use of Menger's imbedding theorems for $(k + 1)$ -tuples, ($1 \leq k \leq n$). A simpler argument by the writer, using a weaker four-point property, is quite independent of those results, and so solved the space problem without any reference to the subset problem [1. pp. 123-128]. (2). The four-point property of Wilson suggests numerous weaker properties which have been investigated by the writer and others [2]. *This paper is concerned with an investigation of weak four-point properties that arise in the metric study of elliptic spaces.*

2. First metric axiomatization of elliptic space. Metric postulates for spherical and hyperbolic spaces, arising from their metric characterizations with respect to the class of semimetric spaces, were established by the writer in 1935 and 1937, respectively.¹ But the numerous metric abnormalities of elliptic space rendered its investigation (in the purely metric manner imposed by the program) a more difficult matter, and it was not until 1946 that the first set of metric postulates for finite and infinite dimensional elliptic spaces was obtained [3]. Chief among the metric features of elliptic space that make inapplicable the methods used in the metric characterizations of other classical spaces are the following.

(1) *Distinction between congruence and superposability.* Defining a *motion* as a congruent mapping of a space onto itself, two subsets are called *superposable* provided there is a motion that maps one onto the other. In contrast to the other classical spaces, two subsets of elliptic space may be congruent without being superposable.

(2) *Distinction between "contained in" and "congruently contained in".*

¹ See [1].

In any of the classical spaces other than the elliptic, a subset that is congruent with a subset of a subspace is actually contained in a subspace of the same dimension. This is not the case in elliptic space.

(3) *Dependence not a congruence invariant.* An m -tuple of a space is usually called *dependent* when it is contained in an $(m - 2)$ -dimensional subspace. With this convention, a dependent m -tuple of elliptic space may be congruent with one that is not dependent.

(4) *Non-linearity of the equidistant locus.* The locus of points of the elliptic plane that are equidistant from two distinct points consists of *two* mutually perpendicular elliptic lines, and hence no subset contained in two such lines forms a metric basis.

(5) *Cardinality of the maximal equilateral set.* The elliptic plane contains six points with all fifteen distances equal. No equilateral septuple exists in the plane or in *elliptic three-space*. The cardinality of the maximal equilateral subset of elliptic n -space is not known for $n > 3$.

The following set of metric postulates for elliptic space (with positive space constant r) was established in [3]. Let E_r denote a distance space containing at least two points.

POSTULATE I. E_r is *semimetric*.

POSTULATE II. E_r is *metrically convex* (that is, if $a, c \in E_r$, $a \neq c$, E_r contains a point b such that $a \neq b \neq c$ and $ab + bc = ac$).

The point c is said to be between a and c , and the relation is symbolized by writing abc .

POSTULATE III. The diameter of E_r is at most $\pi r/2$.

POSTULATE IV. E_r is *metrically complete*.

POSTULATE V. If $p, q \in E_r$, $pq \neq \pi r/2$, then E_r contains points p^* , q^* such that pqp^* , qpq^* subsist, and $pp^* = qq^* = \pi r/2$.

Two points with distance $\pi r/2$ are called *diametral*. If $p \in E_r$, p^* or $d(p)$ will denote a diametral point of p ; that is, $pp^* = pd(p) = \pi r/2$.

DEFINITION. Three points of E_r (not necessarily pairwise distinct) are *LINEAR* provided the sum of two of the three distances they determine equals the third.

If $p_1, p_2, p_3 \in E_r$ let Δ^* denote the determinant $|\varepsilon_{ij} \cos(p_i p_j/r)|$, ($i, j = 1, 2, 3$), where every ε_{ij} is 1, except that $\varepsilon_{23} = \varepsilon_{32} = -1$.

A symmetric matrix (ε_{ij}) , $\varepsilon_{ij} = \varepsilon_{ji} = \pm 1$, $\varepsilon_{ii} = 1$, ($i, j = 1, 2, \dots, m$) is called an EPSILON MATRIX.

POSTULATE VI. *Let p_0, p_1, \dots, p_4 be any five pairwise distinct points of E_r with (i) two triples linear, and (ii) the determinant Δ^* of three of the points (one of which is common to the two linear triples) negative. Then an epsilon matrix (ϵ_{ij}) , $(i, j = 0, 1, \dots, 4)$ exists such that all principal minors of the determinant $|\epsilon_{ij} \cos(p_i p_j / r)|$, $(i, j = 0, 1, \dots, 4)$ are non-negative.*

These postulates insure that all subspaces of E_r (properly defined) of finite or infinite dimensions are elliptic (that is, congruent with the classical elliptic spaces with space constant r).

To axiomatize elliptic n -space $E_{n,r}$, for a given positive integer n , it suffices to adjoin the following (local) postulate.

POSTULATE VII. *The integer n is the smallest for which a point q_0 of E_r and a spherical neighborhood $U(q_0)$ exist such that each $n + 2$ points p_0, p_1, \dots, p_{n+1} of $U(q_0)$ have the property that if there is an epsilon matrix (ϵ'_{ij}) such that no principal minor of the determinant $|\epsilon'_{ij} \cos(p_i p_j / r)|$, $(i, j = 0, 1, \dots, n + 1)$ is negative, then an epsilon matrix (ϵ_{ij}) exists such that no principal minor of $|\epsilon_{ij} \cos(p_i p_j / r)|$, $(i, j = 0, 1, \dots, n + 1)$ is negative, and the determinant vanishes.*

Interpreted geometrically, Postulate VI asserts that each *quintuple* (of a prescribed subclass of the class of all those quintuples of E_r containing two linear triples) is congruently imbeddable in an elliptic space with space constant r . The condition $\Delta^* < 0$ means that the perimeter of the three points for which it is formed is *less than* πr and imparts a local nature to the postulate. It is observed, moreover, that the specific (elliptic) character of the space defined by Postulates I–VI is determined by Postulate VI alone. In view of the discussion above of four-point properties, it is natural to seek to replace the five-point property expressed in Postulate VI by simpler four-point properties. The suggestion to do so, made in the concluding section of [3], was acted upon in the (unpublished) Missouri doctoral dissertation of J. D. Hankins (supervised by the writer) which provides the basis for the present contribution [5].

3. Classes of quadruples and corresponding four-point properties. The following seven classes of semimetric quadruples of pairwise distinct points play a role in what follows.

A semimetric quadruple p_1, p_2, p_3, p_4 belongs to class

$\{Q_1\}$ if and only if it contains a linear triple,

$\{Q_2\}$ if and only if $p_2 p_3 p_4$ subsists and $p_2 p_3 = p_3 p_4$,

$\{Q_3\}$ if and only if $p_2 p_3 p_4$ subsists, $p_2 p_3 = p_3 p_4$, and the perimeter of

every three of the fourpoints is less than $\pi r + \varepsilon$, where r and ε are arbitrarily chosen positive constants,

$\{Q_4\}$ if and only if $p_2p_3p_4$ subsists and $p_1p_2 = p_1p_4$,

$\{Q_5\}$ if and only if $p_2p_3p_4$ subsists, $p_2p_3 = p_3p_4$, and $p_1p_2 = p_1p_4$,

$\{Q_6\}$ if and only if $p_2p_3p_4$ subsists, $p_2p_3 = 2p_3p_4$, and $p_1p_2 = p_1p_3$,

$\{Q_7\}$ if and only if the quadruple contains two linear triples.

Clearly $\{Q_i\} \subset \{Q_1\}$, ($i = 2, 3, \dots, 7$) and $\{Q_3\} \subset \{Q_2\}$.

DEFINITION. A semimetric space has the elliptic WEAK, FEEBLE, ε -FEEBLE, ISOSCELES WEAK, ISOSCELES FEEBLE, EXTERNAL ISOSCELES FEEBLE four-point property if every quadruple of its points of class $\{Q_1\}$, $\{Q_2\}$, \dots , $\{Q_6\}$, respectively, is congruently imbeddable in an elliptic space with space constant r . The space has the ELLIPTIC STRONG TWO-TRIPLE PROPERTY if each of its quadruples of class $\{Q_7\}$ is congruently imbeddable in an elliptic line.

The writer has established elsewhere the following imbedding theorem.²

THEOREM 3.1. A semimetric m -tuple p_1, p_2, \dots, p_m is congruently imbeddable in elliptic n -space $E_{n,r}$ if and only if (i) $p_i p_j \leq \pi r/2$, ($i, j = 1, 2, \dots, m$), and (ii) there exists an epsilon matrix (ε_{ij}) , ($i, j = 1, 2, \dots, m$), such that the determinant $|\varepsilon_{ij} \cos(p_i p_j)/r|$, ($i, j = 1, 2, \dots, m$), has rank not exceeding $n + 1$, with all non-vanishing principal minors positive.

With the aid of this theorem (a) conditions for the congruent imbedding in elliptic space of each quadruple of the classes $\{Q_1\}$, $\{Q_2\}$, \dots , $\{Q_7\}$ are expressed in terms of the six distances determined by the quadruple, and (b) if a quadruple of class $\{Q_i\}$, ($i = 1, 2, \dots, 6$), is congruently imbeddable in $E_{n,r}$, then it is congruently imbeddable in $E_{2,r}$.

DEFINITION. A Σ_r space is any space for which Postulates I-V are valid.

The following sections investigate Σ_r spaces that have one or more of the four-point properties defined above.

4. Spaces Σ_r with the elliptic weak four-point property. Let $\Sigma_r(w)$ denote a Σ_r space with the elliptic weak four-point property. It is proved in [3] that the weak four-point property is possessed by spaces in which Postulates I-VI are valid; that is, in the presence of Postulates I-V, Postulate VI implies the weak four-point property. This section is

² See [1], p. 208.

devoted to showing that in the same environment, the weak four-point property implies Postulate VI.

The following three theorems were either established in [3] by using the weak four-point property (instead of Postulate VI) or their proofs are immediate.

THEOREM 4.1. *Each $\Sigma_r(w)$ space is metric and every triple of points is congruently imbeddable in $E_{2,r}$.*

THEOREM 4.2. *Two distinct non-diametral points p, q of a $\Sigma_r(w)$ space are endpoints of a unique metric segment (denoted by $\text{seg}(p, q)$).*

COROLLARY. *If $p, q \in \Sigma_r(w)$, $p'q' \in E_{n,r}$, $pq = p'q' \neq \pi r/2$, there exists a unique extension of the congruence $p, q \approx p', q'$ to the congruence $\text{seg}(p, q) \approx \text{seg}(p', q')$.³*

THEOREM 4.3. *If $p, q \in E_r(w)$, $(0 < pq < \pi r/2)$ there is exactly one point p^* of $E_r(w)$ such that pqp^* subsists and $pp^* = \pi r/2$.*

Now if $p, q \in E_r(w)$, $(0 < pq < \pi r/2)$ and p^*, q^* are the unique points diametral to p, q , respectively, with pqp^* and qpq^* subsisting, the unique metric segments $\text{seg}(p, q)$, $\text{seg}(q, p^*)$, $\text{seg}(p^*, q^*)$, $\text{seg}(q^*, p)$ have pairwise at most endpoints in common and it follows that the two metric segments

$$\text{seg}(p, q, p^*) = \text{seg}(p, q) + \text{seg}(q, p^*),$$

$$\text{seg}(p, q^*, p^*) = \text{seg}(p, q^*) + \text{seg}(q^*, p^*),$$

have only p, p^* in common.

DEFINITION. *If $p, q \in E_r(w)$, $(0 < pq < \pi r/2)$, then $\text{seg}(p, q, p^*) + \text{seg}(p, q^*, p^*)$ is called a one-dimensional subspace $E_r^1(p, q)$ of $E_r(w)$, with base points p, q , where pqp^* and qpq^* subsist.*

THEOREM 4.4. *A one-dimensional subspace E_r^1 of $E_r(w)$ is congruent with the elliptic line $E_{1,r}$.*

PROOF. $E_r^1 = \text{seg}(p, q, p^*) + \text{seg}(p, q^*, p^*)$, where p, q are base points of E_r^1 . It follows from the weak four-point property that points a, b, a^*, b^* of an elliptic line $E_{1,r}$ exist such that $p, q, p^*, q^* \approx a, b, a^*, b^*$,

³ The notation $p_1, p_2, \dots, p_k \approx q_1, q_2, \dots, q_k$ signifies that $p_i q_j = q_i q_j$, ($i, j = 1, 2, \dots, k$). The symbol " \approx " is read, "is (are) congruent to".

and the two congruences

$$(*) \quad \text{seg}(p, q, p^*) \approx \text{seg}(a, b, a^*),$$

$$(**) \quad \text{seg}(p, q^*, p^*) \approx \text{seg}(a, b^*, a^*),$$

map $E_r^1(p, q)$ onto $E_{1,r}(a, b)$. To show the mapping is a congruence it is clear that only the two following cases need be examined in detail.

Case 1. $x \in \text{seg}(p^*, q^*), p^* \neq x \neq q^*, y \in \text{seg}(q, p^*), q \neq y \neq p^*$. From qyp^* and qp^*q^* follows yp^*q^* , and in a similar manner yp^*x subsists. Hence $xy = xp^* + p^*y$, and letting x', y' correspond to x, y by the congruences (**), (*), respectively, the same considerations establish $x'y' = x'a^* + a^*y'$. Since $xp^* = x'a^*$ and $p^*y = a^*y'$, then $xy = x'y'$.⁴

Case 2. $x \in \text{seg}(q, p^*), q \neq x \neq p^*, y \in \text{seg}(p, q^*), p \neq y \neq q^*$. Since $qx p^*$, qp^*q^* imply xp^*q^* , and pyq^* , pq^*p^* imply yq^*p^* , the quadruple x, p^*, q^*, y contains two linear triples and hence points x'', y'', p'', q'' of $E_{1,r}$ exist such that $x'', y'', p'', q'' \approx x, y, p^*, q^*$. Since $x', a^*, b^* \approx x, p^*, q^* \approx x'', p'', q''$, a motion G of $E_{1,r}$ onto itself exists with $G(x'', y'', p'', q'') = (x', y', a^*, b^*)$. But $\bar{y}a^* = y''p'' = yp^* = y'a^*$, and $\bar{y}b^* = y''q'' = yq^* = y'b^*$. It follows that $\bar{y} = y'$ (since $a^*b^* \neq \pi r/2$) and so $xy = x'y'$.⁵

LEMMA 4.1. If $s, t \in E_r^1$ ($0 < st < \pi r/2$), then $\text{seg}(s, t) \subset E_r^1$.

The proof may be taken from [3].

LEMMA 4.2. Any pair of distinct points of $E_r(w)$ is contained in a unique subspace E_r^1 .

PROOF. If the pair is non-diametral, the result is proved as in [3]. Let p, p^* denote a diametral point pair of $E_r(w)$ and suppose $q \in E_r(w)$ with $pq p^*$. The unique subspace $E_r^1(p, q)$ contains p, p^* , and by Theorem 4.4, $E_r^1(p, q) \approx E_{1,r}(p', q')$. Let E^* denote any one-dimensional subspace of $E_r(w)$ containing p and p^* , and suppose $x \in E^*, x \neq p, p^*, q$. Since there are two linear triples in the quadruple p, q, x, p^* , then $p, q, x, p^* \approx p'', q'', x'', d(p'')$ of $E_{1,r}(p', q')$, where $p''d(p'') = \pi r/2$. A motion G exists such that $G(p'', q'', x'', d(p'')) = (p', q', \bar{x}, d(p'))$, and one of the relations $p'\bar{x}q', q'\bar{x}d(p'), d(p')\bar{x}d(q'), d(q')\bar{x}p'$ subsists, or \bar{x} coincides with one of the points $p', q', d(p'), d(q')$. But then x satisfies the corresponding relation in the unprimed letters, and Lemma 3.1 yields $E^* \subset E_r^1(p, q)$. Interchanging the roles of E^* and E_r^1 gives $E_r^1(p, q) \subset E^*$.

⁴ Obvious modifications of the argument are used in case $x=q^*, y=q$, etc.

⁵ No difficulties are encountered when x, y are not interior points of the segments from which they are chosen.

LEMMA 4.3. *Two congruent triples p_1, p_2, p_3 and p_1', p_2', p_3' of $E_{2,r}$ are superposable if (i) one of the distances $p_i p_j (i, j = 1, 2, 3)$ equals $\pi r/2$, or (ii) $\Delta^*(p_1, p_2, p_3) < 0$.*

PROOF. The proof is given in [3].

THEOREM 4.5. *Let p_1, p_2, p_3 be three pairwise distinct points of $E_r(w)$ with $\Delta^*(p_1, p_2, p_3) < 0$, and p_1', p_2', p_3' points of $E_{2,r}$ with $p_1, p_2, p_3 \approx p_1', p_2', p_3'$. The congruences*

$$(1) \quad E_r^1(p_1, p_2) \approx E_{1,r}(p_1', p_2'),$$

$$(2) \quad E_r^1(p_1, p_3) \approx E_{1,r}(p_1', p_3'),$$

determine uniquely the congruence,

$$E_r^1(p_1, p_2) + E_r^1(p_1, p_3) \approx E_{1,r}(p_1', p_2') + E_{1,r}(p_1', p_3').$$

PROOF. Since p_1, p_2, p_3 are congruently imbeddable in $E_{2,r}$, and $\Delta^*(p_1, p_2, p_3) < 0$, it follows that $\Delta(p_1, p_2, p_3) = |\cos(p_i p_j)/r|$, ($i, j = 1, 2, 3$) is non-negative, and no one of the distances $p_i p_j (i, j = 1, 2, 3)$ is $\pi r/2$. Hence p_1, p_2 and p_1, p_3 are base points of one-dimensional subspaces $E_r^1(p_1, p_2)$ and $E_r^1(p_1, p_3)$, respectively. The congruences (1), (2) in which p_i and $p_i' (i = 1, 2, 3)$ are corresponding points are unique. If $\Delta(p_1, p_2, p_3) = 0$, then $p_3 \in E_r^1(p_1, p_2)$ and so $E_r^1(p_1, p_2)$ and $E_r^1(p_1, p_3)$ coincide. Similarly, $E_{1,r}(p_1', p_2') = E_{1,r}(p_1', p_3')$ and the theorem follows from Theorem 4.4.

If, now, $\Delta(p_1, p_2, p_3) = \Delta(p_1', p_2', p_3') > 0$, then (since $\Delta^*(p_1, p_2, p_3) < 0$) the points p_1', p_2', p_3' neither lie on an elliptic line, nor are they congruent with points of a line, and so p_1, p_2, p_3 are not contained in any E_r^1 of $E_r(w)$. The congruences (1), (2) give a mapping of $E_r^1(p_1, p_2) + E_r^1(p_1, p_3)$ onto $E_{1,r}(p_1', p_2') + E_{1,r}(p_1', p_3')$. To prove the mapping a congruence, suppose $x \in E_r^1(p_1, p_2)$, $y \in E_r^1(p_1, p_3)$, and let x', y' denote their corresponding points by congruences (1), (2), respectively.

CASE I. $x \in \text{seg}(p_1, p_2)$, $p_1 \neq x \neq p_2$, $y \in \text{seg}(p_1, p_3)$. The possibilities $y = p_1$, $y = p_3$ offer no difficulties. Supposing that $p_1 y p_3$ holds, then points p_1'', y'', p_2'', p_3'' of $E_{2,r}$ exist with $p_1, y, p_2, p_3 \approx p_1'', y'', p_2'', p_3''$, and since $p_1', p_2', p_3' \approx p_1, p_2, p_3 \approx p_1'', p_2'', p_3''$, with $\Delta^*(p_1, p_2, p_3) < 0$ a motion G of $E_{2,r}$ exists such that $G(p_1'', y'', p_2'', p_3'') = p_1', \bar{y}, p_2', p_3'$. It is easily seen that $\bar{y} = y'$ and hence $p_2' y' = p_2 y$.

Now from $p_1 x p_2$ follows the existence of points such that $p_1'', p_2'', x'', y'' \approx p_1, p_2, x, y$, where the first quadruple is in $E_{2,r}$, and p_1'', p_2'', y''

are not necessarily those points (with the same notation) considered in the preceding paragraph. From $\Delta^*(p_1, p_2, p_3) < 0$ follows

$$\pi r > p_1 p_2 + p_2 p_3 + p_3 p_1 = p_1 p_2 + p_2 p_3 + p_3 y + y p_1 \geq p_1 p_2 + p_2 y + y p_1,$$

and so $\Delta^*(p_1, p_2, y) < 0$. This permits applying to the quadruple p_1, p_2, x, y the argument applied above to p_1, p_2, p_3, y , and the congruence $p_1, p_2, x, y \approx p_1', p_2', x', y'$ is obtained, yielding $xy = x'y'$.

CASE II. $x \in \text{seg}(p_2, d_1(p_1))$, $p_2 \neq x \neq d_1(p_1)$, $y \in \text{seg}(p_1, p_3, d_2(p_1))$, $p_1 \neq y \neq d_2(p_1)$, where $d_1(p_1)$, $d_2(p_1)$ denote points of $E_r^1(p_1, p_2)$, $E_r^1(p_1, p_3)$, respectively, that are diametral to p_1 .

Let $\{q_j\}$ be a point sequence of $E_r^1(p_1, p_2)$ with the limit p_1 , and $p_1 q_j p_2$ ($j = 1, 2, \dots$). An index m exists such that $q_m p_1 + p_1 y + y q_m < \pi r$. For setting $k = \pi r/2 - p_1 y > 0$, and selecting m so that $q_m p_1 < k/2$ gives $q_m p_1 + p_1 y + y q_m \leq 2(q_m p_1 + p_1 y) < \pi r - k$. It follows that $\Delta^*(p_1, q_m, y) < 0$.

The quadruple p_1, p_3, q_m, y contains the linear triple p_1, p_3, y , and consequently $p_1, p_3, q_m, y \approx p_1'', p_3'', q_m'', y''$, with the latter quadruple in $E_{2,r}$. By Case I, $p_1, p_3, q_m \approx p_1', p_3', q_m'$, and since $\Delta^*(p_1, q_m, y) < 0$ for each point y of $\text{seg}(p_1, p_3, d_2(p_1))$, then $\Delta^*(p_1'', p_3'', q_m'') = \Delta^*(p_1, p_3, q_m) < 0$, and a motion of $E_{2,r}$ sending p_1'', p_3'', q_m'' into p_1', p_3', q_m' , respectively, gives $p_1, p_3, q_m, y \approx p_1', p_3', q_m', \bar{y}$. The linearity of p_1, p_3, y implies that of p_1', p_3', \bar{y} and p_1', p_3', y' . Consequently $p_1', p_3', y \approx p_1, p_3, y \approx p_1', p_3', y'$ implies $\bar{y} = y'$ and $q_m y = q_m' y'$.

Turning now to the quadruple p_1, q_m, x, y , the linearity of p_1, q_m, x together with the relations $p_1, q_m, y \approx p_1', q_m', y'$, $\Delta^*(p_1, q_m, y) < 0$, permits applying the above procedure to obtain $xy = x'y'$.

The various cases arising from x and/or y coinciding with one of the points $p_1, p_2, p_3, d_1(p_1), d_2(p_1)$ are all easily handled, and we may conclude that

$$(3) \quad \text{seg}(p_1, p_2, d_1(p_1)) + \text{seg}(p_1, p_3, d_2(p_1)) \approx \text{seg}(p_1', p_2', d_1(p_1')) + \text{seg}(p_1', p_3', d_2(p_1')).$$

CASE III. $x \in E_r^1(p_1, p_2)$, $y \in E_r^1(p_1, p_3)$, with $p_2 p_1 x$, $p_3 p_1 y$ subsisting, and $\Delta^*(p_1, p_2, y) < 0$, $\Delta^*(p_1, x, y) < 0$.

The method used in Case I is readily applied to yield $p_3 x = p_3' x'$ and $p_2 y = p_2' y'$. Now $p_1, p_2, x, y \approx p_1'', p_2'', x'', y''$, points of $E_{2,r}$. The relations $p_1, p_2, y \approx p_1', p_2', y'$, $\Delta^*(p_1', p_2', y') = \Delta^*(p_1, p_2, y) < 0$, $p_2, p_1 x$, $p_1 p_2 \neq \pi r/2$ yield $p_1'', p_2'', x'', y'' \approx p_1', p_2', x', y'$, and so $xy = x'y'$.

Let o_1, o_2 denote points that satisfy the conditions imposed above on x, y , respectively. Using those points in place of p_2, p_3 , and proceeding as in Case II yields

$$(4) \quad \text{seg}(p_1, o_1, d_1(p_1)) + \text{seg}(p_1, o_2, d_2(p_1)) \approx \\ \text{seg}(p_1', o_1', d_1(p_1')) + \text{seg}(p_1', o_2', d_2(p_1')).$$

ASSERTION. $\text{seg}(p_1, o_1, d_1(p_1)) + \text{seg}(p_1, p_3, d_2(p_1)) \approx \text{seg}(p_1', o_1', d_1(p_1')) + \text{seg}(p_1', p_3', d_2(p_1'))$.

PROOF. Suppose $x \in \text{seg}(o_1, d_1(p_1))$, and $y \in \text{seg}(p_1, p_3, d_2(p_1))$. It is easily seen that $\text{seg}(p_1, o_1)$ contains an interior point q , arbitrarily close to p_1 , such that $qp_1 + p_1y + yq < \pi r$, and $qp_1 + p_1p_3 + p_3q < \pi r$.

By (4), $p_1, q, o_2 \approx p_1', q', o_2'$, and $\Delta^*(p_1', q', o_2') < 0$ follows from $\Delta^*(p_1, o_1, o_2) < 0$ and p_1qo_1 . The familiar procedure now yields $p_1, p_3, q, o_2 \approx p_1', p_3', q', o_2'$, points of $E_{2,r}$. Similarly, it is shown that $p_1, p_3, q, y \approx p_1', p_3', q', y'$. Finally, $p_1, q, x, y \approx p_1', q'', x'', y''$, points of $E_{2,r}$, since xqp_1 holds, and from $p_1, q, y \approx p_1', q', y'$, $\Delta^*(p_1', q', y') < 0$, $p_1'q'x'$, $p_1'q' \neq \pi r/2$, it follows that $p_1, q, x, y \approx p_1', q', x', y'$, and $xy = x'y'$.

Cases not explicitly treated above are either trivial or are handled in a similar manner.

THEOREM 4.6. *Postulate VI is valid in $E_r(w)$.*

PROOF. Let p_0, p_1, p_2, p_3, p_4 be any five pairwise distinct points of $E_r(w)$ with $\Delta^*(p_0, p_1, p_2) < 0$, and each of the triples p_0, p_1, p_3 and p_0, p_2, p_4 linear. Then by Theorem 4.5 the sum $E_r^1(p_0, p_1) + E_r^1(p_0, p_2)$ is congruently imbeddable in $E_{2,r}$, and since p_3, p_4 are elements of the first and second summand, respectively, the five points p_0, p_1, \dots, p_4 are congruently imbeddable in $E_{2,r}$.

It follows from Theorem 3.1 that the quintuple has the property stated in the conclusion of Postulate VI.

THEOREM 4.7. *Postulates I, II, III, IV, V, VI_w, VII are metric postulates for elliptic n -space, where Postulate VI_w postulates the elliptic weak four-point property.⁶*

5. Metric spaces with the elliptic feeble four-point property. The objective of this section is to show that if Postulate I be strengthened to

⁶ Postulate VI_w may be formulated to make Postulate III unnecessary.

require *metricity*, then the class of quadruples assumed congruently imbeddable in $E_{2,r}$ may be restricted to the proper subclass $\{Q_2\}$ of class $\{Q_1\}$. Whether this restriction may be made *without* strengthening Postulate II is an open question. Let $E_r(f)$ denote a *metric* space with the elliptic feeble fourpoint property in which Postulates II–V are valid.

The following theorems are easily established.

THEOREM 5.1. *Each point triple of $E_r(f)$ is congruently imbeddable in $E_{2,r}$.*

THEOREM 5.2. *Two distinct non-diametral points of $E_r(f)$ are joined by exactly one metric segment.*

This follows from (1) the existence of at least one metric segment joining each two distinct points of any complete, metrically convex, metric space, (2) the uniqueness of midpoints for nondiametral pointpairs of $E_r(f)$ (a consequence of the feeble fourpoint property, since such points are unique in $E_{2,r}$), and (3) the fact that each segment of $E_r(f)$ is the closure of the dyadically rational points of the segment.

COROLLARY. *The congruence $p, q \approx p', q', 0 < pq < \pi r/2, (p', q' \in E_{2,r})$ has a unique extension to the congruence $\text{seg}(p, q) \approx \text{seg}(p', q')$.*

REMARK. *There is exactly one $\text{seg}(p, q, p^*)$, with $p, q, p^* \in E_r(f)$ and pqp^* subsisting.*

LEMMA 5.1. *If $p, s, m, d(s) \in E_r(f)$ such that $sd(s) = \pi r/2$, m is a midpoint of $s, d(s)$ (that is, $sm = md(s) = (\frac{1}{2})sd(s)$) and $pd(s) < \pi r/2$, then points $p', s', m', d(s')$ of $E_{2,r}$ exist such that $(p, s) + \text{seg}(m, d(s)) \approx (p', s') + \text{seg}(m', d(s'))$, where $(p, s), (p', s')$ denote the sets consisting of the points exhibited.*

PROOF. If m_1 denotes the unique midpoint of $p, d(s)$, the feeble fourpoint property gives $m_1, s, m, d(s) \approx m_1', s', m', d(s')$, with the latter points in $E_{2,r}$. Similarly, $p, m_1, d(s), m \approx p'', m_1', d(s''), m''$, points of $E_{2,r}$, and since $\Delta^*(m, m_1, d(s)) < 0$, a motion of $E_{2,r}$ yields $p, m_1, d(s), m \approx p', m_1', d(s'), m'$. The theorem is proved by showing that the mapping

$$p \leftrightarrow p', \quad s \leftrightarrow s', \quad \text{seg}(m, d(s)) \approx \text{seg}(m', d(s'))$$

is a congruence.

If $x \in \text{seg}(m, d(s))$ then $sx = sm + mx = s'm' + m'x' = s'x'$. Since $p, m_1, d(s), s \approx p'', m_1'', d(s''), s'', m_1, s, d(s) \approx m_1', s', d(s')$, and $sd(s) = \pi r/2$, Lemma 4.3 yields $p, m_1, d(s), s \approx p^*, m_1', d(s'), s'$. Then p^*, m_1' ,

$d(s') \approx p'$, m_1' , $d(s')$, $m_1'd(s') \neq \pi r/2$, and $p^* \in E_{1,r}(m_1', d(s'))$, (since $p m_1 d(s)$ holds), imply $p^* = p'$, and so p , m_1 , $d(s)$, $s \approx p'$, m_1' , $d(s')$, s' .

It suffices now to show that $p x = p' x'$, for x an interior point of $\text{seg}(m, d(s))$. If m_2 denotes the unique midpoint of $m, d(s)$, the above procedure is applied to obtain p , m_1 , m_2 , $d(s) \approx p'$, m_1' , m_2' , $d(s')$, where m_2' is the midpoint of $m', d(s')$. A continuation of the process yields p , m_1 , q , $d(s) \approx p'$, m_1' , q' , $d(s')$, for each dyadically rational point q of $\text{seg}(m, d(s))$, (that is, for each point q of $\text{seg}(m, d(s))$ such that $m q = \gamma \cdot m d(s)$, where γ denotes any dyadically rational number). Then $p q = p' q'$, and since the set of all the points q is dense in the segment, continuity of the metric gives $p x = p' x'$.

LEMMA 5.2. *Let p , s , m , $d(s)$ denote pairwise distinct points of $E_r(f)$ such that (1) $s d(s) = \pi r/2$, (2) m is a midpoint of $s, d(s)$, and (3) $x \in \text{seg}(s, m, d(s))$ implies $p x < \pi r/2$. Points p' , s' , m' , $d(s')$ of $E_{2,r}$ exist such that*

$$(p) + \text{seg}(s, m, d(s)) \approx (p') + \text{seg}(s', m', d(s')).$$

PROOF. By Lemma 5.1, points p' , s' , m' , $d(s')$ of $E_{2,r}$ exist such that $\text{seg}(s, m, d(s)) \approx \text{seg}(s', m', d(s'))$, $p s = p' s'$, and $p x = p' x'$ if $x \in \text{seg}(m, d(s))$. The lemma is proved by showing that the mapping defined by

$$p \leftrightarrow p', \quad \text{seg}(s, m, d(s)) \approx \text{seg}(s', m', d(s'))$$

is a congruence.

It suffices to prove that $p y = p' y'$, $y \in \text{seg}(s, m)$, $s \neq y \neq m$. Now $\text{seg}(s, m, d(s))$ contains a point \bar{x} such that $p x \leq p \bar{x} < \pi r/2$ for every point x of that segment. Let $\alpha = \pi r/2 - p \bar{x}$, and subdivide $\text{seg}(m, y)$ into $n + 1$ equal subsegments by means of points $q_0 = m, q_1, q_2, \dots, q_{n+1} = y$, such that $q_i q_{i+1} < \alpha$, and $q_{i-1} q_i q_{i+1}$ subsists, ($i = 1, 2, \dots, n$). If $t \in \text{seg}(m, d(s))$ with $m t = m q_1$, then $\Delta^*(p, m, t) < 0$, and $p, m, t \approx p', m', t'$ by the preceding lemma. It follows that $p, m, t, q_1 \approx p', m', t', q_1'$, and so $p q_1 = p' q_1'$. Since $\Delta^*(p, m, q_1) < 0$, and $p, m, q_1 \approx p', m', q_1'$, the above procedure yields $p q_2 = p' q_2'$, and repeated application of the process gives $p y = p q_{n+1} = p' q'_{n+1}$.

LEMMA 5.3. *If s , $d(s)$, p , q denote four points of $E_r(f)$ with $s q d(s)$, $s d(s) = \pi r/2$, and $p x_1 = p x_2 = \pi r/2$ for two distinct points x_1, x_2 of $\text{seg}(s, q, d(s))$, then $(p) + \text{seg}(s, q, d(s)) \approx (p') + \text{seg}(s', q', d(s'))$, and p' is the pole of $\text{seg}(s', q', d(s'))$.*

The proof, based upon the superposability of any two congruent triples of $E_{2,r}$ with a pair of corresponding distances equal to $\pi r/2$, offers no difficulty.

LEMMA 5.4. *If $s, m, d(s), p$ are four pairwise distinct points of $E_r(f)$ such that (1) $sd(s) = \pi r/2$, (2) m is a midpoint of $s, d(s)$, (3) $px_0 = \pi r/2$ for exactly one point x_0 of $\text{seg}(s, m, d(s))$, then $(p) + \text{seg}(s, m, d(s))$ is congruently imbeddable in $E_{2,r}$.*

PROOF. If x_0 is an endpoint of $\text{seg}(s, m, d(s))$, the labelling may be selected so that $x_0 = s$. Then by Lemma 5.1 points $p', s', m', d(s')$ of $E_{2,r}$ exist such that

$$(p, s) + \text{seg}(m, d(s)) \approx (p', s') + \text{seg}(m', d(s')).$$

Let y denote any interior point of $\text{seg}(s, m)$. The procedure of Lemma 5.2 may be applied to show that $py = p'y'$, and continuity of the metric gives $ps = p's'$. The same argument applies in case $x_0 \neq s, m, d(s)$, and the remaining case ($x_0 = m$) is immediate from Lemma 5.1.

The preceding lemmas establish the following theorem.

THEOREM 5.3. *Any subset of $E_r(f)$ consisting of the union of a point and a segment joining two diametral points is congruently imbeddable in $E_{2,r}$.*

Let I_m denote the strengthened form of Postulate I.

THEOREM 5.4. *Postulates $I_m, II, III, IV, V, VI_f, VII$ are metric postulates for elliptic n -space, where VI_f postulates the elliptic feeble four-point property.*

PROOF. It suffices to show that VI_f implies VI_w . If $p, q, s, t \in E_r(f)$ with qst subsisting, then $qt = \pi r/2$ implies $(p) + \text{seg}(q, s, t)$ congruently imbeddable in $E_{2,r}$ (Theorem 5.3), and hence so are p, q, s, t . In case $qt \neq \pi r/2$, then by Postulate V, $E_r(f)$ contains a point $d(q)$ such that $qtd(q)$ and $qd(q) = \pi r/2$. Now $s \in \text{seg}(q, t) \subset \text{seg}(q, t, d(q))$, and hence the congruent imbedding in $E_{2,r}$ of $(p) + \text{seg}(q, t, d(q))$ implies that p, q, s, t are also imbeddable in $E_{2,r}$.

6. Metric spaces with the elliptic ϵ -feeble four-point property. This section is devoted to showing that the class of quadruples assumed imbeddable can be restricted to class $\{Q_3\}$, a proper subclass of $\{Q_2\}$. Let $E_r(\epsilon - f)$ denote a space satisfying Postulates $I_m - V$, with every quadruple of class $\{Q_3\}$ congruently imbeddable in $E_{2,r}$.

It is easily seen that Theorem 5.2, together with the Corollary and Remark following it, are valid under the weaker assumption made in this section.

THEOREM 6.1. *The union of any segment of $E_r(\varepsilon - f)$ joining a diametral pointpair, and any point of the space is congruently imbeddable in $E_{2,r}$.*

PROOF. Let $p, s, d(s)$ be points of $E_r(\varepsilon - f)$ with $sd(s) = \pi r/2$, and let $X = [x \in \text{seg}(s, d(s)) \mid px = \pi r/2]$. Clearly, X is a closed set.

Case I. X is null. Then $\text{seg}(s, d(s))$ admits a partition into equal non-overlapping subsegments so small in length that the perimeter of each triple of points contained in any quadruple formed by p and three adjacent points effecting the partition is less than πr . An argument similar to that used in the proof of Lemma 5.2 may be applied.

Case II. $X = (s)$ or $X = (d(s))$. Select the labelling so that $X = (s)$, and let t be an interior point of $\text{seg}(s, d(s))$. It is easily seen that points $p', t', d(s')$ of $E_{2,r}$ exist such that

$$(p) + \text{seg}(t, d(s)) \approx (p') + \text{seg}(t', d(s')),$$

with $p, t, ds \approx p', t', d(s')$. Extend $\text{seg}(t', d(s'))$ to s' so that $s't'd(s')$ and $s'd(s') = \pi r/2$ subsist. The mapping defined by

$$p \leftrightarrow p', \quad \text{seg}(s, t, d(s)) \approx \text{seg}(s', t', d(s'))$$

is easily seen to be a congruence.

Case III. $X = (t)$, $t \in \text{seg}(s, d(s))$, $s \neq t \neq d(s)$. If $u, v \in \text{seg}(s, d(s))$ with sut and $tvd(s)$ subsisting and $ut = tv < \varepsilon/2$, the perimeter of every triple in the quadruple p, u, t, v is less than $\pi r + \varepsilon$. Then points p', u', v', t' of $E_{2,r}$ exist with $p, u, t, v \approx p', u', t', v'$. We have $\text{seg}(s, t, d(s)) \approx \text{seg}(s', t', d(s'))$, where the latter segment of $E_{2,r}$ contains $\text{seg}(u', t', v')$.

The congruence

$$(p) + \text{seg}(u, t, v) \approx (p') + \text{seg}(u', t', v')$$

is easily established and its extension to

$$(p) + \text{seg}(s, t, d(s)) \approx (p') + \text{seg}(s', t', d(s'))$$

is proved by the method of Lemma 5.2.

Case IV. X contains at least two points. Then every point of $\text{seg}(s, d(s))$ belongs to X and the desired conclusion is immediate.

THEOREM 6.2. *Postulates I_m , II, III, IV, V, $VI(\varepsilon - f)$, VII are metric postulates for elliptic n -space, where $VI(\varepsilon - f)$ postulates the elliptic ε -feeble four-point property.*

PROOF. It is clear from Theorem 6.1 that the space has the feeble four-point property, and so the theorem follows from Theorem 5.4.

It is worth remarking that the argument used in establishing the basic Theorem 6.1 requires ε to be positive. It seems likely, however, that the theorem is valid if ε be replaced by zero; that is, if the congruent imbedding in $E_{2,r}$ of all quadruples p, q, s, t with $qs = st = (\frac{1}{2})qt$ and the perimeter of each triple of points less than πr , be assumed.

7. Metric spaces with the elliptic isosceles weak four-point property and the elliptic strong two-triple property. This section is concerned with spaces for which Postulates I_m –V are valid and such that all quadruples of classes $\{Q_4\}$ and $\{Q_7\}$ are congruently imbeddable in $E_{2,r}$. Denote the space E_r (i.w.t.t.).

The imbeddability in $E_{2,r}$ of quadruples of class $\{Q_7\}$ suffices to establish the following theorems and remarks.⁷

THEOREM 7.1. *Each two distinct non-diametral points of E_r (i.w.t.t.) are joined by a unique metric segment.*

THEOREM 7.2. *If $p, q \in E_r$ (i.w.t.t.), $0 < pq < \pi r/2$, there is exactly one point $d(p)$ of the space with $pqd(p)$ and $pd(p) = \pi r/2$.*

REMARK 1. If $pqd(p)$, $pd(p) = \pi r/2$, there is a unique $\text{seg}(p, q, d(p))$.

REMARK 2. The relations $pqd(p)$, $qp d(g)$, $pd(p) = qd(q) = \pi r/2$, imply $pd(q)d(p)$.

THEOREM 7.3. *A one-dimensional subspace of E_r (i.w.t.t.) is congruent with $E_{1,r}$.*

On the other hand, the proof of the following basic theorem makes no direct use of the congruent imbedding of quadruples of class $\{Q_7\}$, but uses only the imbedding of quadruples of class $\{Q_4\}$.

THEOREM 7.4. *Let p be any point and E_r^1 any one-dimensional subspace of E_r (i.w.t.t.). The $E_{2,r}$ contains a point p' and a line $E_{1,r}$ such that*

$$(p) + E_r^1 \approx (p') + E_{1,r}.$$

⁷ See [1], pp. 217–220.

PROOF. The theorem follows from Theorem 7.3 in case $p \in E_r^1$, and is obviously valid if $px = \pi r/2$ for every point x of E_r^1 . It may be assumed, therefore, that if f denotes a foot of p on E_r^1 , then $0 < pf < \pi r/2$. Let a, b be points of E_r^1 such that a/b and $af = fb = \pi r/4$.

ASSERTION. *The point f is the only foot of p on $\text{seg}(a, f, b)$.*

If there were two additional feet f_1, f_2 of p on $\text{seg}(a, f, b)$, then the $E_{2,r}$ contains points f_1', f_2', f', p' with $f_1, f_2, f, p \approx f_1', f_2', f', p'$. But $p'f_1' = p'f_2' = p'f'$ and the linearity of f_1', f_2', f' imply $p'f = p'f' = \pi r/2$, contrary to the above.

Suppose, now, that f_1 is a foot of p on $\text{seg}(a, f, b)$, $f \neq f_1$, and denote by g the midpoint of f, f_1 . From the congruent imbedding of p, f, g, f_1 in $E_{2,r}$ follows $pq = \pi r/2$. Assume the labelling so that gf_1b or $f_1 = b$. If x is interior to $\text{seg}(a, f)$, then $pf < px \leq pq$, and so a point y of $\text{seg}(f, g)$ exists such that $px = py$. Similarly, a point z of $\text{seg}(g, f_1)$ exists such that $px = py = pz$. Imbedding p, x, y, z in $E_{2,r}$ yields $px = py = pz = \pi r/2$, and imbedding p, f, x, y in $E_{2,r}$ gives $pf = \pi r/2$. Hence the Assertion is proved.

Select $\text{seg}(x, y)$ on E_r^1 so that f is its midpoint, $xy < \pi r/2$, and $px + xy + py < \pi r$. If $px = py$, let the labels q, s replace x, y , respectively, while in the contrary case, label so that $px < py$. In the latter event, a point z of $\text{seg}(f, y)$ exists such that $pz = px, z \neq f$. Now let q be the label of x , and s the label of z . Then $q, f, s, p \approx q', f', s', p'$, with the "primed" points in $E_{2,r}$ and $q'f's'$ holding. The proof is completed by showing that the correspondence defined by

$$p \leftrightarrow p', \quad E_r^1(q, f, s) \approx E_{1,r}(q', f', s')$$

is a congruence, where the congruence exhibited in the correspondence is an extension of $g, f, s \approx g', f', s'$. If $x \in E_r^1(q, f, s)$, its correspondent in $E_{1,r}(q', f', s')$ will be denoted by "primes".

If $g \in \text{seg}(q, f, s)$ it is easily seen that $p, q, g, s \approx p', q', g', s'$, and consequently

$$(*) \quad (p) + \text{seg}(q, f, s) \approx (p') + \text{seg}(q', f', s').$$

It follows that f' is the foot of p' on $\text{seg}(q', f', s')$, and $q'f' = qf = fs = f's'$. In the triangles (p', f', s') and (p', f', q') the angles $\angle p'f's'$ and $\angle p'f'q'$ are right angles. Let x, y be points of E_r^1 such that fqx subsists, $qx < \min(aq, fq)$, xfy holds, with $fy \leq fx$, and $px = py$.

Then $p, x, y, f \approx p'', x'', y'', f''$, with the latter quadruple in $E_{2,r}$.

ASSERTION. *The point f'' is the foot of p'' on $E_{1,r}(x'', f'', y'')$.*

From $p''x'' = px = py = p''y''$, the foot of p'' on $E_{1,r}(x'', f'', y'')$ is either the midpoint m'' of $\text{seg}(x'', y'')$ or its diametral point $d(m'')$ on that line. The latter alternative is easily seen to be impossible.

Now the midpoint m of $\text{seg}(x, y)$ is a point of $\text{seg}(q, s)$, and so $pm < \pi r/2$. From the congruent imbedding in $E_{2,r}$ of p, x, m, y it is seen that m'' and f'' coincide, and the Assertion is established.

The equalities $xf = x''f'' = f''y'' = fy$ follow, and letting x', y' denote the points on $E_{1,r}(q', f', s')$ corresponding to x'', y'' , respectively, yields $x''f'' = x'f' = f'y' = f''y''$, and $p''f'' = pf = p'f'$. The two elliptic right triangles (p'', f'', y'') , (p', f', y') are congruent, and so $py = p''y'' = p'y' = p'x'$. Also $p''y'' = py = px$, and consequently $px = p'x'$. Thus congruence (*) can be extended to

$$(p) + \text{seg}(x, f, y) \approx (p') + \text{seg}(x', f', y'),$$

and since $pz < \pi r/2$, $z \in \text{seg}(x, y)$, repetition of the procedure yields

$$(**) \quad (p) + \text{seg}(a, f, b) \approx (p') + \text{seg}(a', f', b').$$

Let $d(f)$ denote the point of $E_{r^1}(a, f, b)$ that is diametral to f , and suppose g is an interior point of $\text{seg}(a, d(f))$. Since agb and $pa = pb$ hold, $p, a, g, b \approx p'', a'', g'', b''$, points of $E_{2,r}$, and $p'', a'', b'' \approx p, a, b \approx p', a', b'$. Since $a'b' = \pi r/2$, the two triples are superposable, and a motion gives $p, a, g, b \approx p', a', g^*, b'$, with g^* on $E_{1,r}(a', f', b')$ and $a, b, g^* \approx a', b', g'$. Then either $g' = g^*$ or g^* is the reflection of g' in a' . The latter case is easily seen to be impossible. Hence $p, a, g, b \approx p', a', g', b'$ and $pg = p'g'$.

In a similar manner it is seen that $p'v = p'v'$, for v' an interior element of $\text{seg}(b, d(f))$, and the theorem is proved.

It follows at once that E_r (i.w.t.t.) has the elliptic weak four-point property and the following axiomatization results.

THEOREM 7.5. *Postulates I_m , II, III, IV, V, VI (i.w.t.t.), VII are metric postulates for elliptic n -space, where Postulate VI (i.w.t.t.) asserts that every quadruple of classes $\{Q_4\}$ and $\{Q_7\}$ are congruently imbeddable in $E_{2,r}$.*

8. Metric spaces with quadruples of classes $\{Q_5\}$, $\{Q_6\}$, $\{Q_7\}$ imbeddable in $E_{2,r}$. By virtue of the strong two-triple property (the imbedding of quadruples of class $\{Q_7\}$), Theorems 7.1, 7.2, 7.3 of the preceding section are valid here. Let E_r^* denote any metric space satisfying the demands of

Postulates II–V, and such that each quadruple of its points belonging to $\{Q_5\}$, $\{Q_6\}$, or $\{Q_7\}$ is congruently imbeddable in $E_{2,r}$.

THEOREM 8.1. *The set sum of any point and any one-dimensional subspace of E_r^* is congruently imbeddable in $E_{2,r}$.*

PROOF. It suffices to consider the case of $p \in E_r^*$, $E_r^1 \subset E_r^*$, f a foot of p on E_r^1 , and $0 < pf < \pi r/2$. It is not difficult to show that (1) f is unique and (2) E_r^1 contains at most one point g with $pg = \pi r/2$. If such a point g exists, and f^* , g^* denote those points of E_r^1 diametral to f , g , respectively, let x_1 , x_2 be points of $\text{seg}(f, g)$, $\text{seg}(f, g^*, f^*)$, respectively, such that $fx_1 < fx_2$. It may be shown that $px_1 < px_2$.

Let a , b be the two midpoints of f, f^* on E_r^1 , and suppose $g \in \text{seg}(f, a, f^*)$. Choose points x , y of E_r^1 so that $x \in \text{seg}(a, f)$, $y \in \text{seg}(f, b)$, $2xy < fg$, $px = py$, and $\angle^*(p, x, y) < 0$. Assume that the midpoint m of x , y is distinct from f . Points s , t of E_r^1 exist such that $s \in \text{seg}(x, f)$, $t \in \text{seg}(f, y)$, $ps = pt$ and $st = 2 \cdot ty$. Then $p, s, t, y \approx p'', s'', t'', y''$, points of $E_{2,r}$. Let z be a point of E_r^1 such that $ts = 2 \cdot sz$. Then it is easily seen that $p, s, t, z \approx p'', s'', t'', z''$, and $p''z'' = p''y''$. Since $p''y'' = py = px$ and $px = pz$, with z interior to $\text{seg}(f, g)$, it follows that $x = z$. Hence $xs = ty$, and m is the midpoint of s , t . Repeating this procedure a finite number of times establishes m as the midpoint of a pair of points for which it is *not a betweenpoint*. As a result of this contradiction, we conclude that pairs of points in $\text{seg}(x, y)$ having f as their midpoint are equidistant from p .

Select points a', f', b', p' of $E_{2,r}$ such that $E_r^1(a, f, b) \approx E_{1,r}(a', f', b')$ is an extension of $a, f, b \approx a', f', b'$, and f' is the foot of p' on $E_{1,r}(a', f', b')$, with $p'f' = pf$. The proof is completed by showing that the mapping

$$p \leftrightarrow p', \quad E_r^1(a, f, b) \approx E_{1,r}(a', f', b')$$

is a congruence.

THEOREM 8.2. *Postulates I_m , II, III, IV, V, VI*, VII are metric postulates for elliptic n -space, where Postulate VI* asserts the congruent imbedding in $E_{2,r}$ of all point quadruples of classes $\{Q_5\}$, $\{Q_6\}$, $\{Q_7\}$.*

9. A fundamental unsolved problem. It is known that every semimetric space is congruently imbeddable in the $E_{2,r}$ whenever each 7 of its points are so imbeddable [7] ⁸. This is stated by saying that the elliptic plane has congruence indices $\{7, 0\}$ with respect to the class of semimetric

⁸ See also [4], in which congruence indices $\{8, 0\}$ are proved.

spaces. Since the $E_{2,r}$ contains an equilateral sextuple, the congruence indices $\{7, 0\}$ are the best (that is, for no integers m, k ($m < 7$) are indices $\{m, k\}$ valid). This result, together with Theorem 3.1, completely solves the congruent imbedding problem for $E_{2,r}$ and hence provides a metric axiomatization for the class of subsets of $E_{2,r}$. Metric postulates for the $E_{2,r}$ itself are obtained by adjoining any metric properties that serve to distinguish the elliptic plane among its subsets, though, as observed earlier in this paper, this approach to the metric characterization of a space is likely to result in a redundant set of postulates.

Perhaps the most important unsolved problem suggested by this manner of studying elliptic geometry is the determination of congruence indices for $E_{n,r}$ when $n > 2$. The problem for the $E_{2,r}$ is quite difficult, and the methods employed there seem incapable of extension even to $E_{3,r}$. Apparently an entirely different approach is needed. At the present time not even any preliminary results concerning the problem for general dimension have been obtained.

Bibliography

- [1] BLUMENTHAL, L. M., *Theory and applications of distance geometry*. The Clarendon Press, Oxford 1953, XI + 347 pp.
- [2] —, *An extension of a theorem of Jordan and von Neumann*. Pacific Journal of Mathematics, vol. 5 (1955), pp. 161–167.
- [3] —, *Metric characterization of elliptic space*. Transactions American Mathematical Society, vol. 59 (1946), pp. 381–400.
- [4] —, and KELLY, L. M., *New metric-theoretic properties of elliptic space*. Revista de la Universidad Nacional de Tucumán, vol. 7 (1949), pp. 81–107.
- [5] HANKINS, J. D., *Metric characterizations of elliptic n -space*. University of Missouri doctoral dissertation, 1954.
- [6] MENGER, K., *Untersuchungen über allgemeine Metrik*. Mathematische Annalen, vol. 100 (1928), pp. 75–163.
- [7] SEIDEL, J., *De Congruentie-orde van het elliptische vlak*. Thesis, University of Leiden, 1948, iv + 71 pp.
- [8] WILSON, W. A., *A relation between metric and euclidean spaces*. American Journal of Mathematics, vol. 54 (1932), pp. 505–517.

AXIOMS FOR GEODESICS AND THEIR IMPLICATIONS

HERBERT BUSEMANN

University of Southern California, Los Angeles, California, U.S.A.

The foundations of geometry are principally concerned with elementary geometry and in particular with the role of continuity. Although our intuition relies on continuity, very large sections of euclidean and non-euclidean geometry prove valid without continuity hypotheses.

This lecture deals with the foundations of metric differential geometry. Continuity is taken for granted and the interest centers on the question *to which extent differentiability is necessary*. In order to delineate the subject more clearly we emphasize that we do not mean results like that of Wald [15], who characterizes Riemannian surfaces with a continuous Gauss curvature among metric spaces, because his point is not the weakening or omission of differentiability properties but their replacement by other limit processes.

Two major advances have been made in the indicated direction, one — our principal subject — is due to the author and concerns, roughly, *the intrinsic geometry in the large of not necessarily Riemannian spaces*; most of this theory can be found in the book [3]. The second is the work of A. D. Alexandrov and deals with *surfaces either in E^3 or with an abstract intrinsic Riemannian metric*. Much of this material is found in his book [1], a brief survey in [6]. In both theories the tools created in order to do without differentiability assumptions proved in many instances far superior to the classical, in fact they yield a number of results which remain inaccessible to the traditional methods, even when smoothness is granted.

It also appears that a frequently followed procedure, which works, as it were, from the top down by reducing differentiability hypotheses in existing proofs, has, in general, very little chance of producing final results.

The axioms for a G-space, [3, Chapter I]. Since we are interested in metric differential geometry our first axiom is:

I. *The space is metric.*

We call the space R , denote points by small Roman letters, the distance from x to y by xy . Since the concept of metric space has been generalized in various ways we mention explicitly that we require the standard properties: $xx = 0$, $xy = yx > 0$ if $x \neq y$, and the triangle inequality $xy + yz \geq xz$. But large parts of the theory hold without the symmetry condition $xy = yx$.

The relations $x \neq y$, $y \neq z$ and $xy + yz = xz$ will be written briefly as (xyz) . A set M in R is bounded if $xy < \beta$ for a suitable β and all x, y in M .

Our second axiom is the validity of the Bolzano Weierstrass theorem:

II. *A bounded infinite set has an accumulation point.*

In conjunction with the following axioms, II entails that the space is complete and behaves in all essential respects like a finite-dimensional space. Whether the axioms actually imply finite dimensionality is an open question.

The third axiom guarantees that the metric is intrinsic. It was introduced by Menger as convexity of a metric space:

III. *If $x \neq z$, then a point y with (xyz) exists.*

It follows from I, II, III that any two points x, y can be connected by a segment $T(x, y)$, i.e., a set isometric to an interval $[\alpha, \beta]$ of the real t -axis. $T(x, y)$ can therefore be represented in the form $z(t)$, $\alpha \leq t \leq \beta = \alpha + xy$ with

$$(1) \quad z(t_1)z(t_2) = |t_1 - t_2|,$$

and $z(\alpha) = x$, $z(\beta) = y$.

These axioms are satisfied, for example, by a closed convex subset of a euclidean space. However, we aim at geometries which cannot be extended without increasing the dimension. Obviously some form of prolongability is necessary. Requiring that any segment can be prolonged would be too strong, it would eliminate even the ordinary spherical metric. If $S(p, \rho)$ denotes the set of points x with $px < \rho$, we postulate:

IV. *Every point p has a neighborhood $S(p, \rho_p)$, $\rho_p > 0$, such that for any two distinct points x, y in $S(p, \rho_p)$ a point z with (xyz) exists.*

The generality of the function ρ_p is deceiving; the axiom furnishes a function $\rho(p) > 0$ satisfying IV and the Lipschitz condition $|\rho(p) - \rho(q)| \leq pq$.

The four axioms yield geodesics. A *geodesic* is a locally isometric image of the real t -axis, precisely: it can be represented in the form $z(t)$,

$-\infty < t < \infty$, and there is a positive function $\varepsilon(t)$ such that (1) holds for $|t_i - t| \leq \varepsilon(t)$, $i = 1, 2$. Thus, the geodesics on an ordinary cylinder are either entire helices, entire straight lines or circles traversed infinitely often.

Geodesics exist in the following sense: a function $z(t)$ satisfying (1) in an interval $\alpha \leq t \leq \beta$, $\alpha < \beta$, can be extended to all real t , so that it represents a geodesic. This is the analogue to the indefinite continuation of a line element into a geodesic in the classical case.

Axioms I to IV contain no uniqueness properties. In an (x_1, x_2) -plane metrized by $xy = |x_1 - y_1| + |x_2 - y_2|$ any curve $z(s) = (z_1(s), z_2(s))$, $a \leq s \leq b$, for which both $z_1(s)$ and $z_2(s)$ are monotone is a segment from $z(a)$ to $z(b)$. We observe that in the classical case a segment can be prolonged by a given amount in at most one way and therefore postulate:

V. If (xyz_1) , (xyz_2) and $yz_1 = yz_2$, then $z_1 = z_2$.

The five axioms guarantee that the above extension $z(t)$ of a segment to a geodesic is *unique*. Moreover, if (xyz) then $T(x, y)$ and $T(y, z)$ are unique (because two different $T(y, z)$ would yield two different prolongations of $T(x, y)$). In particular, $T(x, y)$ is unique for $x, y \in S(p, \rho_p)$, so that the *local uniqueness of the shortest connection*, which is so important for many investigations in differential geometry, need not be explicitly stipulated.

The spaces satisfying the five axioms are called *G-spaces*, the *G* alluding to geodesic.

There are two particularly simple types of geodesics, namely those which satisfy (1) for arbitrary t_1, t_2 and are therefore isometric to the entire real axis; they are called *straight lines*. The others are the so-called *great circles* which are isometric to ordinary circles. A representation $z(t)$ of a *great circle* of length β is characterized by

$$z(t_1)z(t_2) = \min_{|\nu| \ 0, 1, 2, \dots} |t_1 - t_2 + \nu\beta|.$$

The cylinder shows that straight lines, great circles and geodesics which are neither may occur in one space. When IV holds in the large, or z with (xyz) exists for any two distinct points x, y , then all geodesics are straight lines (and conversely), and the space is called *straight*.

The lowest dimensional *G-spaces* are uninteresting. A 0-dimensional space is obviously a point and a one-dimensional *G-space* is a straight line or a great circle. The two-dimensional *G-spaces* can be proved to be topological manifolds; the corresponding problem for higher dimensions is open.

It is important to notice that the *axioms comprise the Finsler spaces*, where the line element has the form $ds = f(x_1, \dots, x_n; dx_1, \dots, dx_n) = f(x, dx)$, and $f(x, dx)$ satisfies certain standard conditions (see [3, Section 15]) but need not be quadratic or Riemannian $ds^2 = \sum g_{ik}(x) dx_i dx_k$. The analytical methods often become highly involved for Finsler spaces. This explains why the limitation of the hypotheses inherent to the axiomatic approach leads in this case to improved methods, which have the additional appeal of effecting a *synthesis of differential geometry, topology, the calculus of variations, the foundations of geometry, and convex body theory*.

Spaces with negative curvature, [3, Chapter V]. It is impossible to outline the whole theory of G -spaces in the space available here. We therefore restrict ourselves to giving a few typical results and discuss in greater detail only the theory of parallels which is more closely related to the remaining geometric topics of this symposium.

Hadamard discovered in [8] that the *surfaces with negative curvature* have many beautiful properties. For Riemann spaces his results were extended by others in various directions. The very concept of curvature seems to imply notions of differentiability. However, each of the following two properties proves in the Riemannian case to be equivalent to non-positive (negative) curvature:

For each point p there is a positive δ_p , $0 < \delta_p \leq \rho_p$ such that

- (a) *if a, b, c lie in $S(p, \delta_p)$ but not on a segment, and b', c' are the mid-points of $T(a, b)$ and $T(a, c)$, then $2b'c' \leq 2bc$ ($2b'c' < 2bc$);*
- (b) *if $C(T, \epsilon)$ denotes the set $xT \leq \epsilon$ where T is a segment and $C(T, \epsilon) < S(p, \delta_p)$ then $C(T, \epsilon)$ is convex (strictly convex).*

Convexity is defined in the usual way by means of the $T(x, y)$.

In Finsler spaces, hence also in G -spaces, (a) is stronger than (b). The geometry discovered by Hilbert [9, Anhang I] (which corresponds to the Klein model of hyperbolic geometry with a convex curve replacing the ellipse as absolute locus) furnishes an example where (b) holds but not (a). In fact, a Hilbert geometry satisfying (a) is hyperbolic (see Kelly and Strauss [11]). The condition (b) was introduced by Pedersen [12].

Practically all of Hadamard's and the later results on Riemann spaces with non-positive or negative curvature hold for G -spaces with the property of domain invariance satisfying (a), and many are valid when (b) holds. In particular, (b) implies that the universal covering space of the given

space R is straight. Consequently, for two given points of R , the geodesic connection is unique within a given homotopy class. Moreover, when the $C(T, \epsilon)$ are strictly convex, then there is at most one closed geodesic within a given class of freely homotopic curves; if compact, the space cannot have an abelian fundamental group and possesses only a finite number of motions (isometries of the space on itself). The latter result is, for Riemann spaces, contained in Bochner [2] and for the general case in [7]. The theory of parallels (see below) requires (a) instead of (b).

Characterizations of the elementary spaces, [3, Chapter VI]. For brevity we call the euclidean, hyperbolic, and spherical spaces *elementary*. The *bisector* $B(a, a')$ of two distinct points a, a' in a G -space is the locus of the points x equidistant from a and a' ; that is, $ax = a'x$. The *elementary spaces*, with the exception of the 1-sphere, are characterized by the fact that their bisectors contain with any two points x, y at least one segment $T(x, y)$. The principal theorem is the following local version:

(2) Let $0 < \delta \leq \rho_p$ and assume that for any five distinct points a, a', b, c, x in $S(p, \delta)$ the relations $ab = a'b$, $ac = a'c$ and (bxc) entail $ax = a'x$. Then $S(p, \delta)$ is isometric to an open sphere of radius δ in an elementary space.

The hypothesis means, of course, that $b, c \in B(a, a')$ implies $x \in B(a, a')$. Whereas the proofs for the results on spaces with non-positive curvature are not essentially longer than they would be under differentiability hypotheses, such hypotheses would very materially shorten the long proof of (2) in [3, Sections 46, 47].

Various well-known theorems are corollaries of (2); examples are the following global and local answers to the *Helmholtz-Lie Problem*:

If for two given isometric triples a_1, a_2, a_3 and a_1', a_2', a_3' (i.e., $a_1a_3 = a_1'a_3'$) of a G -space R a motion of R exists taking a_i into a_i' , $i = 1, 2, 3$, then R is elementary.

If every point of a G -space R has a neighborhood $S(p, \delta)$, $0 < \delta < \rho_p$, such that for any four points a_1, a_2, a_1', a_2' in $S(p, \delta)$ with $pa_1 = pa_1'$, $pa_2 = pa_2'$ and $a_1a_2 = a_1'a_2'$ a motion of $S(p, \delta)$ exists which takes a_i into a_i' , $i = 1, 2$, then the universal covering space of R is elementary.

By using deeper results from the modern theory of topological and Lie groups, Wang [16] and Tits [14] succeeded to determine all spaces with the property that for any two pairs a_1, a_2 and a_1', a_2' with $a_1a_2 = a_1'a_2'$ a

motion exists taking a_i into a'_i , $i = 1, 2$. If the space has an odd dimension then it is either elementary or elliptic; for even dimensions greater than 2 there are other solutions.

Inverse problems in the large for surfaces, [3, Sections 11, 33], [4], [13]. In inverse problems of the calculus of variations one gives a set of curves and asks whether they occur as the extremals of a variational problem. The local inverse problem in two dimensions was solved by Darboux, but his method provides no answers in the large. Inverse problems in the large cannot be treated by one method, they differ depending on the topological structure of the surface, and are inaccessible to the traditional approach.

Three of these problems have been solved with the present methods. We mention first that a *G-space*, in which the geodesic through two (distinct) points is unique, is either straight, or all its geodesics are great circles of the same length β (see [3, Theorem (31, 2)]). In the latter case we say that the space is of the *elliptic type*. If its dimension exceeds 1, then it has a two-sheeted universal covering space which shares with the sphere the properties that all geodesics are great circles of the same length 2β , and that all geodesics that pass through a given point meet again at a second point.

A two-dimensional *G-space* R , in which the geodesic through two points is unique, is therefore either homeomorphic to the euclidean plane E^2 or to the projective plane P^2 . If in the case of E^2 a euclidean metric $e(x, y)$ is introduced, then the geodesics of R form a system N of curves which have, in terms of $e(x, y)$, the following two properties:

- 1) Each curve in N is representable in the form $z(t)$, $-\infty < t < \infty$, with $z(t_1) \neq z(t_2)$ for $t_1 \neq t_2$ and $e(z(0)), z(t) \rightarrow \infty$ for $|t| \rightarrow \infty$.
- 2) There is exactly one curve of N through two distinct points.

The answer to the corresponding inverse problem is:

If a system N of curves in E^2 with the properties 1), 2) is given, then the plane can be remetrized as a G -space with the curves in N as geodesics.

It will become clear from examples later that the problem of determining *all* metrics with the curves in N as geodesics has too many solutions to be interesting.

The inverse problem P^2 was solved by Skornjakov [13], a simpler proof is found in [4]:

In P^2 let a system N' of curves homeomorphic to a circle be given and such that there is exactly one curve of N' through 2 given distinct points. Then P^2 may be metrized as a G -space with the curves in N' as geodesics.

The third problem solved with these methods is that of a *torus with a straight universal covering space*. It differs from the preceding problems in that there are non-obvious necessary conditions. In the plane as the universal covering space of the torus we can introduce an auxiliary euclidean metric $e(x, y)$ and cartesian coordinates x_1, x_2 such that the covering transformations are the translations $T(m_1, m_2)$:

$$x_1' = x_1 + m_1, \quad x_2' = x_2 + m_2, \quad m_1, m_2 \text{ integers.}$$

To the geodesics on the torus there then corresponds in the plane a system N of curves which satisfies the conditions 1), 2) above and in addition the following:

- 3) N goes into itself under the $T(m_1, m_2)$.
- 4) If a curve in N passes through q and $qT(m_1, m_2)$ then it also passes through the points $qT(vm_1, vm_2)$, $|v| = 1, 2, \dots$
- 5) N satisfies the parallel axiom (on its usual form, see below).

Whereas it is not hard to establish 4), the proof of 5) is far from obvious and represents, as far as the author is aware, the only instance in the literature where the validity of the parallel axiom appears as a non-trivial theorem.

If a system N of curves in the plane is given which has the properties 1) to 5), then the plane can be metrized as a G -space with the curves in N as geodesics, where the metric is invariant under the $T(m_1, m_2)$ and thus yields a metrization of the torus.

The curves in N need not satisfy Desargues' Theorem, even when N and the metric arc invariant under $T(m_1, r)$, where r is an arbitrary real number. These facts exhibit very clearly the great generality of Finsler spaces as compared to Riemann spaces: *the only Riemannian metrizations of the torus such that the universal covering space is straight are euclidean* (see E. Hopf [10]).

The theory of parallels, [3, Chapter III]. In the foundations of geometry the congruence axioms, parallel axiom (euclidean or hyperbolic), and the continuity axioms usually appear in this order. The present theory

suggests the study of parallelism on the basis of continuity without congruence or mobility axioms.

In a straight space denote by $G(a, b)$, $a \neq b$, the geodesic, briefly *line*, through a and b and by $G^+(a, b)$ the same line with the orientation in which b follows a . Let G^+ be any oriented line, p any point. Then $G^+(p, x)$ converges to a line A^+ , when x traverses G^+ in the positive direction. The convergence of $G^+(p, x)$ is trivial in the plane, but not in higher dimensions. A^+ is called the *asymptote* to G^+ through p . It is independent of p in the sense that for $q \in A^+$ the line $G^+(q, x)$ also tends to A^+ .

Denote by A^- , G^- the opposite orientations of the lines A , G carrying A^+ , G^+ . If A^+ and A^- are asymptotes to G^+ and G^- respectively, then we call A *parallel* to G . These definitions suggest investigating the following properties:

SYMMETRY: *If A^+ is an asymptote to G^+ , then G^+ is an asymptote to A^+ . If A is parallel to G , then G is parallel to A .*

TRANSITIVITY: *If A^+ is asymptote to B^+ , and B^+ is to C^+ , then so is A^+ to C^+ .*

It is very easily seen that the *transitivity of the asymptote relation implies its symmetry*. The converse holds in the plane, but it is not known whether this extends to higher dimensions.

Even in a plane *the asymptotic relation is not always symmetric*. In an (x, y) -plane let H, H_1 be the branches $x < 0$ of the hyperbolas $xy = -1$ and $xy = 1$ respectively. Let H^-, H_1^- be their orientations corresponding to decreasing x . The system N consists of all curves obtainable by translations from H , of the lines $y = mx + b$, $m < 0$ and the lines $x = \text{const.}$. The system N_1 consists of the curves obtainable by translations from H or H_1 and of the lines $x = \text{const.}$, $y = \text{const.}$ Each of the systems satisfies the conditions 1)–2) above and hence serves as system of geodesics for a G -space.

Denote by Y^+ and Y_{-1}^+ the lines $y = 0$, $y = -1$ with the orientations corresponding to increasing y . In both systems N, N_1 , Y^+ is an asymptote to Y_{-1}^+ and so is Y^- to Y_{-1}^- ; thus, Y is parallel to Y_{-1} . *In N the line Y_{-1}^+ is not the asymptote to Y^+ through $(-1, 1)$, but H^+ is, whereas Y_{-1}^- is an asymptote to Y^- . In the system N_1 neither Y_{-1}^+ is an asymptote to Y^+ nor is Y_{-1}^- to Y^- .*

In the plane the *parallel axiom* in its usual form (namely, *if $p \notin G$ then there is exactly one line A through p which does not intersect G*) is equivalent

to postulating that *for any p and G , if A^+ is the asymptote to G^+ through p , then so is A^- to G^-* . The uniqueness of the non-intersecting line implies symmetry and transitivity.

The usual formulation of the corresponding *hyperbolic axiom* (if $p \notin G$ and A^+ is the asymptote to G^+ through p , then A^- is not an asymptote to G^-) does *not* imply symmetry. The intersections of the curves in the system N just constructed with the domain $x < 0$, $y > -x$ provide an example. This corresponds to the fact that in the foundations of hyperbolic geometry symmetry and transitivity of the asymptote relation are proved with the help of the congruence axioms.

Other questions concern the *distances* from A^+ to G^+ . In any straight space the existence of points $x_\nu \in A^+$ and $y_\nu \in G^+$ which tend on A^+ and G^+ in the positive direction to ∞ and for which $x_\nu y_\nu \rightarrow 0$ is *sufficient* for A^+ and G^+ to be asymptotes to each other. But boundedness of xG , when x traverses a positive subray of A^+ is neither *necessary* nor *sufficient* for A^+ to be an asymptote to at least one orientation of G . In fact, very surprising phenomena occur even for the ordinary lines as geodesics.

Let $g(t)$ be defined and continuous for $t \geq 0$, $g(0) = 0$, $g(t_1) < g(t_2)$ for $t_1 < t_2$ and $g(t) \rightarrow \infty$ for $t \rightarrow \infty$. Put

$$f(x, \alpha) = \text{sign}(x_1 \cos \alpha + x_2 \sin \alpha)g(|x_1 \cos \alpha + x_2 \sin \alpha|).$$

Then the arguments of [3, Section 11] show readily that

$$(3) \quad \rho_g(x, y) = \int_{-\pi/2}^{\pi/2} |f(x, \alpha) - f(y, \alpha)| d\alpha$$

is a metrization of the (x_1, x_2) -plane as a G -space with the lines $ax_1 + bx_2 + c = 0$ as geodesics. $\rho_g(x, y)$ is invariant under $x_1' = x_1 \cos \alpha - x_2 \sin \alpha$, $x_2' = x_1 \sin \alpha + x_2 \cos \alpha$, so that the metric even possesses the rotations about $(0, 0)$. Nevertheless, simple estimates show that for $g(t) = \log(1+t)$ any two parallel lines G_1, G_2 have the property that $\rho_g(x, G_2) \rightarrow 0$ when x traverses G_1 in either direction. For $g(t) = e^t - 1$, we have $\rho(x, G_2) \rightarrow \infty$.

In straight spaces which satisfy the condition (a) for nonpositive curvature, the boundedness of xG when x traverses a positive subray of A^+ is necessary and sufficient for A^+ to be an asymptote to a suitable orientation of G , so that the asymptote relation is transitive.

For the foundations of geometry it is of greater interest to see how *mobility* eliminates the various abnormal occurrences. Assume a plane P which is metrized as a straight space possesses a motion α which, reduced to the straight line G , is a proper translation of G ($z(t)\alpha = z(t+a)$, $a \neq 0$).

Then the asymptote relation is transitive within the families of asymptotes to G^+ and G^- (see [3, Section 32]). The parallels to G are exactly those lines which go into themselves under α . They either cover all of P , or a closed halfplane, or a closed strip which may reduce to G . If p does not lie on a parallel to G , let G_1 be the parallel to G (possibly G itself) closest to p . If x traverses an asymptote A^+ to G_1^+ (or G^+) in the positive direction, then $xG_1 \rightarrow 0$; if x traverses A^+ in the negative direction, then $xG_1 \rightarrow \infty$.

A characterization of the higher dimensional euclidean geometry, [3, Theorem (24.10)]. In the foundations of geometry parallelism for lines in space reduces to that in a plane, because only spaces are considered in which any three points lie in a plane. In straight spaces of higher dimension than two we mean by the *parallel axiom* the following two requirements:

The asymptote relation is symmetric. If A^+ is an asymptote to G^+ then so is A^- to G^- .

The metrics $\rho_g(x, y)$ which can be extended to higher dimensions, show that this parallel axiom and the Theorem of Desargues or the existence of planes do not imply that the space is Minkowskian (finite dimensional linear). Without any postulates regarding the existence of planes the *higher dimensional euclidean geometry is characterized by the above parallel axiom together with the existence and symmetry of perpendicularity in this sense*:

(4) If $p \notin G$, $f \in G$, and $pf = \min_{x \in G} px$, then for every $x \in G$ also $xf = \min_{x \in G(p, f)} xy$.

It is well-known, see [3, p. 104], that there are Minkowski planes which are not euclidean and satisfy (4). Also, (4) is without the parallel axiom a weak condition; it is, for example, satisfied by every simply connected Riemann space with non-positive curvature. This follows from [3, Theorems (20.9) and (36.7)] and the symmetry of perpendicularity in Riemann spaces.

Similarities and differentiability, [5]. It is natural to ask how we can recognize from the behaviour of our finite distances xy whether a G -space possesses differentiability properties in terms of suitable coordinates. The best guide in a statement which is formulable and false without, but correct with differentiability hypotheses.

A *proper similarity* of a G -space R is a mapping α of R on itself such that $x\alpha y\alpha = kxy$ for all $x, y \in R$ where k is a positive constant different from 1. A *proper similarity* has exactly one fixed point f . Because α^{-1} is also a similarity and its factor is k^{-1} , we may assume that $k < 1$. Then $x\alpha^n y\alpha^n = k^n xy$ and $x\alpha^n x\alpha^{n+1} = k^n xx\alpha$ show that $x\alpha^n$ is a Cauchy sequence with a limit f and also that $y\alpha^n \rightarrow f$. It follows readily that the space is straight.

Linear spaces obviously possess similarities with arbitrary factors k , but this does not characterize them among all G -spaces without differentiability hypotheses. For if in (3) we choose $g(t) = t^\beta$, $\beta > 0$, then $\rho_g(\delta x, \delta y) = \delta^\beta \rho(x, y)$, where $\delta x = (\delta x_1, \delta x_2)$, so that for $\delta^\beta = k$ the mapping $x \rightarrow \delta x$ is a similarity with the factor k ; yet the space is linear only for $\beta = 1$.

Differentiability always means a locally nearly linear behaviour. We say that a G -space is *continuously differentiable at p* if for any sequence of triples of distinct points a_ν, b_ν, c_ν which tend to p , and for any points b'_ν, c'_ν with $(a_\nu b'_\nu b_\nu)$, $(a_\nu c'_\nu c_\nu)$ and $a_\nu b'_\nu : a_\nu b_\nu = a_\nu c'_\nu : a_\nu c_\nu = t_\nu$, we have

$$\lim_{\nu \rightarrow \infty} b'_\nu c'_\nu / t_\nu b_\nu c_\nu = 1.$$

(Differentiability would correspond to the special case $a_\nu = p$ and proves insufficient as our example shows.) A G -space is *Minkowskian* if it possesses one *proper similarity* α and is *continuously differentiable at the fixed point of α* .

This form of differentiability is adequate also for the problem originally posed: Let a G -space R be continuously differentiable at p . Put $p_t = p$, and for $q \neq p$, let $(pq_t q)$ and $pq_t : pq = t$. Then for $x, y \in S(p, \rho_p)$, the limit

$$m_p(x, y) = \lim_{t \rightarrow 0} x_t y_t / t$$

exists. Obviously $m_p(p, x) = px$. The metric $m_p(x, y)$ can be extended so as to yield a space satisfying I, II, III, and IV in the strong form that a point z with (xyz) exists for any $x \neq y$. In general V will not hold for this "tangential metric" at p . If V does hold we say — following the terminology of the calculus of variations — that the space is regular at p .

If the space is continuously differentiable and regular at p , then the metric $m_p(x, y)$ is Minkowskian and $xy/m_p(x, y) \rightarrow 1$, for $x \neq y$ and $x \rightarrow p, y \rightarrow p$.

If the space R is a Finsler space with $ds = f(x, dx)$ as above and R is of class C^m , $m > 4$, and f is of class C^{m-1} for $dx \neq 0$, then R will, in $S(p, \rho_p) - p$, be at least of class C^{m-2} , and f of class C^{m-3} in affine coordinates belonging to m_p (normal coordinates). Thus we obtain a complete decision of the problem whether a G -space is a Finsler space of class C^∞ and a partial solution for finite m .

Two dimensional Riemann spaces, [1]. The great variety of metrics satisfying the Axioms I to V indicates that relaxing these axioms essentially without adding others leads to spaces with too little structure for a significant theory. On the other hand there are surfaces in E^3 like polyhedra and general convex surfaces which do not satisfy our axioms, hence still less the assumptions of classical differential geometry, but are geometrically most interesting.

It is the purpose of A. D. Alexandrov's theory to define and study (intrinsically and extrinsically) a class of surfaces narrow enough to encompass deep results and yet wide enough to include, for example, the mentioned surfaces. He assumes that the space R be a two-dimensional manifold, metrized such that the distance of any two points x, y equals the greatest lower bound of the lengths of all curves from x to y . (If II holds, this implies III.) The problem is *how to introduce the Riemannian character of the metric without differentiability*. Alexandrov's principal tool is the (*upper*) angle $\alpha(T, T')$ between two segments T, T' with the same origin z : If $x(t), y(t), t > 0, x(0) = y(0) = z$, represents T and T' , then

$$\alpha(T, T') = \lim_{t \rightarrow 0} \sup_{s \rightarrow 0} \arccos \frac{t^2 + s^2 - [x(t)y(s)]^2}{2ts},$$

where $0 \leq \arccos \leq \pi$. For a geodesic triangle D with sides T, T', T'' we define the *excess* as

$$\varepsilon(D) = \alpha(T, T') + \alpha(T', T'') + \alpha(T'', T) - \pi.$$

The Riemannian character of the metric enters through the requirement that for every compact subset M of R a number $\beta(M)$ exists such that for any finite set of non-overlapping triangles D_1, \dots, D_m , in M

$$(5) \quad \sum_{i=1}^m |\varepsilon(D_i)| < \beta(M).$$

According to Zalgaller [17] it suffices to require $\sum \varepsilon(D_i) < \beta(M)$, in other words, the triangles with negative excess never cause any trouble.

That the condition (5) is really essentially Riemannian follows from the fact that a Minkowski plane does not satisfy it unless it is euclidean.

To these so-called *surfaces with bounded curvature* Alexandrov extended some of the deepest theorems of differential geometry; we mention only Weyl's problem: *Let the sphere or the plane be metrized such that I, II, III and (5) hold. Assume moreover that the excess $\varepsilon(D) \geq 0$ for all small geodesic triangles. Then the metric can be realized by a closed or by a complete open convex surface in E^3 .* Actually the open surfaces were not covered by the classical methods, and the results of Alexandrov on the deformation of convex surfaces surpass by far anything obtainable by the traditional approach.

The theorem of Nash-Kuiper on the C^1 -imbeddability in E^3 of given abstract two-dimensional Riemannian manifolds in the classical sense stresses the significance of these results, because it shows that a reasonable and general class of surfaces in E^3 cannot be defined in terms of differentiability conditions only.

Bibliography

- [1] ALEXANDROW, A. D., *Die innere Geometrie der konvexen Flächen*, Berlin 1955, XVII + 522 pp.
- [2] BOCHNER, S., *Vector fields and Ricci curvature*, Bulletin of the American Mathematical Society, vol. 52 (1946) pp. 776-797.
- [3] BUSEMANN, H., *The geometry of Geodesics*. New York 1955, X + 422 pp.
- [4] ———, *Metrizations of projective spaces*. Proceedings of the American Mathematical Society, vol. 8 (1957) pp. 387-390.
- [5] ———, *Similarities and differentiability*. Tôhoku Mathematical Journal, Sec. Ser., vol. 9 (1957), pp. 56-67.
- [6] ———, *Convex Surfaces*. New York 1958, VII + 194 pp.
- [7] ———, *Spaces with finite groups of motions*. Journal de Mathématiques pures et appliquées, 9th Ser., vol. 37 (1958) pp. 365-373.
- [8] HADAMARD, J., *Les surfaces à courbures opposées et leur lignes géodésiques*. Journal de Mathématiques pures et appliquées, 5th. Ser. vol. 4 (1898), pp. 27-73.
- [9] HILBERT, D., *Grundlagen der Geometrie*. 8th. ed., Stuttgart 1956, VII + 251 pp.
- [10] HOPF, E., *Closed surfaces without conjugate points*. Proceedings of the National Academy of Sciences, vol. 34 (1948) pp. 47-51.
- [11] KELLY, P. J., and STRAUS, E. G., *Curvature in Hilbert geometry*. Pacific Journal of Mathematics, vol. 8 (1958) pp. 119-126.

- [12] PEDERSEN, F. P., *On spaces with negative curvature*. Matematisk Tidsskrift B 1952, pp. 66–89.
- [13] SKORNYAKOV, L. A., *Metritzation of the projective plane in connection with a given system of curves* (Russian). Izvestiya Akademii Nauk SSSR, Seriya Matematicheskaya 19 (1955) pp. 471–482.
- [14] TITS, J., *Sur certaines classes d'espaces homogènes de groupes de Lie*. Académie royale de Belgique, Classe des sciences, Mémoires, Collection in-8°, vol. 39 fasc. 3 (1955), 268 pp.
- [15] WALD, A., *Begründung einer koordinatenlosen Differentialgeometrie der Flächen*. Ergebnisse eines mathematischen Kolloquiums (Wien) Heft 7 (1936), pp. 24–46.
- [16] WANG, H. C., *Two-point homogeneous spaces*. Annals of Mathematics, vol. 55 (1952), pp. 177–191.
- [17] ZALGALI ER, V. A., *On the foundations of the theory of two-dimensional manifolds with bounded curvature* (Russian). Doklady Akademii Nauk SSSR, vol. 108 (1956), pp. 575–576.

AXIOMS FOR INTUITIONISTIC PLANE AFFINE GEOMETRY

A. HEYTING

University of Amsterdam, Amsterdam, Netherlands

1. Introduction. At first sight it may appear that the axiomatic method cannot be used in intuitionistic mathematics, because there are only considered mathematical objects which have been constructed, so that it makes no sense to derive consequences from hypotheses which are not yet realized. Yet the inspection of the methods which are actually used in intuitionistic mathematics, shows us that they are for an important part axiomatic in nature, though the significance of the axiomatic method is perhaps somewhat different from that which it has in classical mathematics.

In principle every theorem can be expressed in the form of an axiomatic theory. Instead of "Every natural number is a product of prime numbers" we can write "Axiom. n is a natural number. Theorem. n is a product of prime numbers.". This way of presentation becomes practicable whenever a great number of theorems contains the same complicated set of hypotheses. Thus, conversely, every axiomatic theory can be read as one general theorem of the form: "Whenever we have constructed a mathematical object M satisfying the axioms A , we can affirm about M the theorems T ."

Of course the content of the theory will be influenced by the intuitionistic point of view; in particular, questions of effective constructibility will be of main importance. In order to give an idea of these differences I shall show the method at work in an example, which I have so chosen that the problem is trivial in classical mathematics, so that the intuitionistic difficulties appear, so to say, in their purest form.

2. The problem. In [1] and [2] I gave a system of axioms for intuitionistic plane projective geometry. Here I wish to give a system of axioms for plane affine geometry which is satisfied by the intuitionistic analytical geometry, and which allows us to construct, by a suitable extension of the plane, a projective plane which satisfies the axioms of [1] and [2]. This problem is easy in the case of desarguesian geometry, because then the

extension can be effected by means of harmonic conjugates. If only the trivial axioms of incidence are assumed, the problem is still easy from the classical point of view, but it presents serious difficulties in intuitionistic mathematics. These difficulties are caused by the fact that not only points at infinity must be adjoined to the affine plane, but also points for which it is unknown whether they are at infinity or not.

3. The axiom system for projective geometry.

FUNDAMENTAL NOTIONS:

Two disjoint sets \mathfrak{P} and \mathfrak{L} ; the elements of \mathfrak{P} are called *points*; those of \mathfrak{L} *lines*.

A relation $\#$, whose domain and range are \mathfrak{P} ; this relation is called *apartness*.

A relation ε , whose domain is \mathfrak{P} and whose range is \mathfrak{L} ; this relation is called *incidence*.

Notation: Capitals in italics denote points; lower case italics denote lines.

Free use is made of such expressions as “a line through a point”; the translation into the incidence – language is left to the reader. Also, line l is sometimes identified with the set of points incident with l , without further explanation. It would be easy to avoid such identifications by a somewhat clumsier presentation. In particular the notation $l \cap m$ is used for the set of points, which are incident with l as well as with m .

Logical signs are used as abbreviations. They must be understood in the intuitionistic sense (see [3] and [4] or [5]).

\rightarrow stands for implication, $\&$ for conjunction, \vee for disjunction, \neg for negation,

$(\forall x)$ is the universal quantifier (for every x),

$(\exists x)$ is the existential quantifier (there exists an x such that).

AXIOMS FOR APARTNESS:

$$S1 \quad A \# B \rightarrow B \# A.$$

$$S2 \quad \neg A \# B \leftrightarrow A = B.$$

$$S3 \quad A \# B \rightarrow (\forall C)(C \# A \vee C \# B).$$

GEOMETRICAL AXIOMS:

P1 $A \neq B \rightarrow (\exists l)(A \in l \ \& \ B \in l)$

P2 $A \neq B \ \& \ A \in l \cap m \ \& \ B \in l \cap m \rightarrow l = m.$

DEFINITION 1. A lies OUTSIDE l ($A \omega l$) if $(\forall B)(B \in l \rightarrow B \neq A).$

DEFINITION 2. l lies APART FROM m ($l \neq m$), if $(\exists A)(A \in l \ \& \ A \omega m).$

P3 $l \neq m \rightarrow (\exists A)(A \in l \cap m).$

P4 $A \neq B \ \& \ A \in l \ \& \ B \in l \ \& \ C \omega l \ \& \ A \in m \ \& \ C \in m \rightarrow B \omega m.$

- P5 (i) There exist two points, A and B , so that $A \neq B$;
 (ij) Every line contains at least three points A, B, C , so that
 $A \neq B, A \neq C$ and $B \neq C$;
 (iij) When l is a line, a point outside l can be found.

DEFINITION 3. If $A \neq B$, then the line l satisfying $A \in l \ \& \ B \in l$ is denoted by AB .

It can be proved from these axioms, that the relation \neq between lines (Definition 2) is an apartness relation; this means that it satisfies axioms S1–S3 for lines instead of points.

4. The axiom system for affine geometry.

FUNDAMENTAL NOTIONS: $\mathfrak{P}, \mathfrak{Q}, \neq, \in$, as in § 3.

AXIOMS FOR APARTNESS: S1–S3, as in § 3.

DEFINITIONS: 1 and 2 as in § 3.

GEOMETRICAL AXIOMS:

A1 $l \neq m, A \omega l \rightarrow (\exists p)(A \in p \ \& \ l \cap p = l \cap m).$

A2 $A \neq B \ \& \ A \in l \cap m \ \& \ B \in l \cap m \rightarrow l = m.$

DEFINITION 4. l INTERSECTS m if $l \neq m \ \& \ (\exists A)(A \in l \cap m).$

A3 l intersects $m \rightarrow (\forall p)((\exists A)(A \in l \cap p) \vee (\exists B)(B \in m \cap p)).$

A4 $A \neq B \ \& \ A \in l \ \& \ B \in l \ \& \ C \omega l \ \& \ A \in m \ \& \ C \in m \rightarrow B \omega m.$

A5 $P \omega l \ \& \ l \cap m = \emptyset \ \& \ P \in m \ \& \ Q \in l \rightarrow Q \omega m.$

DEFINITION 5. l is PARALLEL to m ($l \parallel m$) if $(\forall A)(A \in l \rightarrow A \omega m)$.

Remark: A5 can now be formulated as follows: $l \cap m = \emptyset \& m \neq l \rightarrow l \parallel m$.

A6 $(\forall l)(\exists m)(l \parallel m)$.

A7 (i) *There exists at least one line;*

(ii) *Every line is incident with at least four points every two of which are apart from each other;*

(iii) $A \neq B \rightarrow (\exists l)(A \in l \& B \omega l)$.

(iv) $A \in l \rightarrow (\exists m)(A \in m \& l \neq m)$.

Remarks: (1) If l and m have a common point B , A1 asserts the existence of a line joining A and B (see Th. 1a). On the other hand, if l and m have no common point, it follows from A1 that there is a line through A which does not intersect l ; this is part of the assertion of existence of parallels. Moreover, A1 admits an assertion in the case that it is unknown whether l intersects m . (2) $A2 = P2$. (3) A3 is a strong form of the uniqueness assertion for parallels. (4) A4 is called the triangle axiom.

5. Elementary theorems.

THEOREM 1. *If $l \neq m$, $A \in l \cap m$, $B \in l \cap m$, then $A = B$.*

PROOF: Suppose $A \neq B$; then, by A2, $l = m$, which contradicts $l \neq m$. So $A = B$.

Remark: In the case of Th. 1 we write $l \cap m = A$.

THEOREM 1a. $A \neq B \rightarrow (\exists l)(A \in l \& B \in l)$.

PROOF: By A7(iii) there is a line p so that $A \in p \& B \omega p$.

By A7(iv) there is a line m so that $A \in m \& p \neq m$.

By Th. 1, $p \cap m = A$. By A1 there is a line l so that $B \in l$ and $p \cap l = p \cap m = A$; it follows that $A \in l$.

THEOREM 2. *The relation \neq between lines is an apartness relation, i.e. it possesses the properties (i)–(iii):*

(i) $l \neq m \rightarrow m \neq l$.

(ii) $\neg l \neq m \leftrightarrow l = m$.

(iii) $l \neq m \rightarrow (\forall p)(p \neq l \vee p \neq m)$.

LEMMA 2.1. (iii) holds if l intersects m in S and $S \in p$.

PROOF: Choose points A, B, C so that $A \in l, A \omega m$ (Def. 2); $B \in m, B \neq S; C \in m, C \neq S, C \neq B$. This is possible by A7 (ii), S3. By A4, $B \omega AC$, so $AB \neq AC$. By A3, p has a point in common with AB or with AC ; say $D \in p \cap AB$. $D \neq A \vee D \neq B$. If $D \neq A$, then $D \omega l$ (A4), so $p \neq l$; if $D \neq B$, then $D \omega m$ (A4), so $p \neq m$.

PROOF OF (i): Choose A, B, C so that $A \in l, A \omega m, B \in m, C \in m, B \neq C$. $AB \neq AC$ (A4). $l \neq AB \vee l \neq AC$ (Lemma 2.1). If $l \neq AB$, we choose D on l so that $D \omega AB$; then $B \omega l$ (A4), so $m \neq l$. Similarly, if $l \neq AC$.

LEMMA 2.2. $\neg P \omega l \rightarrow P \in l$.

PROOF: Choose C and m so that $C \omega l, C \in m, P \omega m$; then $l \neq m$. Choose A so that $A \in m, A \omega l$ (Th. 2 (i)). $PA \neq m$. By A3, l has a point in common with PA or with m . If l intersects m in S , we choose B so that $B \in m, B \omega PA$. $PA \neq PB$; l has a point in common with PA or with PB . Say $Q \in l \cap PA$. Now suppose $P \neq Q$; then $P \omega l$, which contradicts the hypothesis, so $\neg P \neq Q$, so $P = Q$ (S2), so $P \in l$.

PROOF OF (ii): $\neg l \neq m \rightarrow l = m$ is an immediate consequence of Lemma 2.2, while $l = m \rightarrow \neg l \neq m$ is a consequence of S2.

PROOF OF (iii): Choose P, Q, R, S so that $P \in l, P \omega m; Q, R, S \in m; Q \neq R \neq S \neq Q$. As in the proof of (i) it can be shown that at least two of the points Q, R, S are outside l ; say $Q, R \omega l$. $PQ \neq PR$, so p has a point in common with PQ or with PR ; say $A \in p \cap PQ$. $A \neq P \vee A \neq Q$. If $A \neq P$, then $A \omega l$, so $p \neq l$. If $A \neq Q$, then $A \omega m$, so $p \neq m$.

THEOREM 3. $P \omega l \rightarrow (\exists m)(P \in m \& m \parallel l)$.

PROOF: Draw a line $p \parallel l$ (A6). By A1, there is a line m through P so that $l \cap m = l \cap p = \emptyset$. By A5, $m \parallel l$.

THEOREM 4. $l \parallel m \& l \parallel n \& m \neq n \rightarrow m \parallel n$.

PROOF: Choose P and Q so that $P \in n, P \omega m, Q \in m$. $PQ \neq m$. By A3, l has a point S in common with PQ . $S \omega n$; by A4, $Q \omega n$. As Q is an arbitrary point on m , we have $m \parallel n$.

6. Projective points.

DEFINITION 6. $l \neq m \rightarrow \mathfrak{P}(l, m) = \{x | l \cap m = l \cap x \vee l \cap m = m \cap x\}$. If $l \neq m$, then $\mathfrak{P}(l, m)$ is a PROJECTIVE POINT (abbreviation: *p. point*).

Remarks: Where $\mathfrak{P}(l, m)$ occurs, it is understood that $l \neq m$. German capitals will be used to denote projective points. The next theorem shows that the notion of a projective point is an extension of that of a line pencil in the usual sense.

THEOREM 5. *If l intersects m in S , then $\mathfrak{P}(l, m)$ is the class of all lines through S . If $l \parallel m$, then $\mathfrak{P}(l, m) = \{x \mid x \parallel l \vee x \parallel m\}$.*

PROOF: The first part of the theorem is obvious. If $l \parallel m$, and $n \in \mathfrak{P}(l, m)$, then $l \cap n = \emptyset$ or $m \cap n = \emptyset$; also $l \neq n$ or $m \neq n$. The only case which needs to be further considered is $l \cap n = \emptyset$ & $m \neq n$. Suppose $S \in m \cap n$; then by A3, l has a point in common with m , in contradiction with $l \parallel m$. Thus $m \cap n = \emptyset$, so $m \parallel n$. Conversely, if $l \parallel m$ and $n \parallel l$, then $l \cap m = l \cap n$, so $n \in \mathfrak{P}(l, m)$.

THEOREM 6. $p, q \in \mathfrak{P}(l, m) \text{ \& } p \neq q \rightarrow \mathfrak{P}(l, m) = \mathfrak{P}(p, q)$.

LEMMA 6.1. $l \neq m \text{ \& } l \neq n \text{ \& } l \cap m = m \cap n \rightarrow l \cap m = l \cap n$.

PROOF: It follows from the hypothesis that $l \cap m \subseteq l \cap n$. Suppose $P \in l \cap n$; then l intersects n , so m has a point in common with l or with n (A3).

Case 1. m intersects l . As $l \cap n = P$ and $l \cap m \subseteq l \cap n$, we have $l \cap m = P = l \cap n$.

Case 2. $Q \in m \cap n$. Then $Q \in l \cap m$, so $Q \in l \cap n$; it follows that $Q = P$ and that $l \cap n \subseteq l \cap m$.

COROLLARY: In the case that $l \neq n$ we have $n \in \mathfrak{P}(l, m) \leftrightarrow l \cap m = l \cap n$.

LEMMA 6.2. $l \neq m \neq n \neq l \text{ \& } n \in \mathfrak{P}(l, m) \rightarrow \mathfrak{P}(l, m) = \mathfrak{P}(l, n)$.

PROOF: By hypothesis and lemma 6.1 we have $l \cap m = l \cap n = m \cap n$. Suppose $p \in \mathfrak{P}(l, m)$. $p \neq l$ or $p \neq m$ & $p \neq n$.

Case 1. $p \neq l$. Then $l \cap m = l \cap p$ (lemma 6.1), so $l \cap n = l \cap p$, so $p \in \mathfrak{P}(l, n)$.

Case 2. $p \neq m$ & $p \neq n$. Now $l \cap m = m \cap p$. $X \in l \cap n \rightarrow X \in l \cap m \rightarrow X \in m \cap p$, so $X \in n \cap p$. It follows that $l \cap n \subseteq n \cap p$. Now suppose $Y \in n \cap p$. By A3, we may distinguish subcases 2a: (m intersects n) and 2b: (m intersects p).

Case 2a. m intersects n in Z . $m \cap n = l \cap m = m \cap p$, so $Z \in n \cap p$. It follows that $Z = Y$, and that $Y \in m \cap n$, so $Y \in l \cap n$.

Case 2b. m intersects p in Z . $Z \in m \cap p = l \cap m = l \cap n$, so $Z \in n \cap p$. It follows that $Y = Z$ and that $Y \in m \cap p = l \cap n$.

In case 2a as well as in case 2b we have proved that $n \cap p \subseteq l \cap n$; thus

in case 2, $l \cap n = n \cap p$, that is $p \in \mathfrak{P}(l, n)$. We have proved that $\mathfrak{P}(l, m) \subseteq \mathfrak{P}(l, n)$. In particular, $m \in \mathfrak{P}(l, n)$; the same proof then gives us $\mathfrak{P}(l, n) \subseteq \mathfrak{P}(l, m)$.

LEMMA 6.3.: $l \# m \& l \# n \& n \in \mathfrak{P}(l, m) \rightarrow \mathfrak{P}(l, m) = \mathfrak{P}(l, n)$.

PROOF: Choose A and B so that $A \in l$, $A \omega m$, $B \in m$, $B \omega l$.

By A3, n intersects $l \vee (\exists C)(C \in n \cap AB)$.

If n intersects l in P , $l \cap m = l \cap n = P$, so $\mathfrak{P}(l, m) = \mathfrak{P}(l, n)$.

If $C \in n \cap AB$, we have $n \# m \vee n \# AB$.

If $n \# m$, we can apply Lemma 6.2.

If $n \# AB$, we choose D so that $D \# A, B, C$ and $D \in AB$. By A1, there is a line p so that $D \in p \& p \in \mathfrak{P}(l, m)$. Now by Lemma 6.2, $\mathfrak{P}(l, m) = \mathfrak{P}(l, p) = \mathfrak{P}(l, n)$.

PROOF OF THEOREM 6: We may suppose that $q \# l$.

Choose n in $\mathfrak{P}(l, m)$ so that $n \# l, q$ (see the proof of lemma 6.3). By Lemma 6.3 we have $\mathfrak{P}(l, m) = \mathfrak{P}(l, n) = \mathfrak{P}(n, q) = \mathfrak{P}(p, q)$.

DEFINITION 7. If l intersects m in S , then $\mathfrak{P}(l, m)$ is a *PROPER p. point* and we write $\mathfrak{P}(l, m) = S$. If $l \parallel m$, then $\mathfrak{P}(l, m)$ is an *IMPROPER p. point*.

Remark: It is by no means true that every p. point is either proper or improper. However, as an immediate consequence of A5, a p. point that cannot be proper, is improper.

DEFINITION 8. l lies OUTSIDE \mathfrak{A} ($l \omega \mathfrak{A}$) if $(\forall p)(p \in \mathfrak{A} \rightarrow p \# l)$.

DEFINITION 9. \mathfrak{A} lies APART from \mathfrak{B} ($\mathfrak{A} \# \mathfrak{B}$) if $(\exists p)(p \in \mathfrak{A} \& p \omega \mathfrak{B})$.

Remark. If \mathfrak{A} is a proper p. point A , then $l \omega \mathfrak{A}$ is equivalent with $A \omega l$. This is easily proved by means of Axiom A4. It follows that, for proper p. points $\mathfrak{A} = A$ and $\mathfrak{B} = B$, $\mathfrak{A} \# \mathfrak{B}$ is equivalent to $A \# B$. Axiom A1 can now be read as follows:

If A is a proper p. point and \mathfrak{B} a p. point so that $A \# \mathfrak{B}$, then there exists a line l so that $A \in l$ and $l \in \mathfrak{B}$. This line will be denoted by $A\mathfrak{B}$. It follows from Th. 8 below that it is unique.

THEOREM 7. The relation $\#$ between projective points is an apartness relation; that is, it possesses the properties (i), (ii), (iii):

- (i) $\mathfrak{A} \# \mathfrak{B} \rightarrow \mathfrak{A} \# \mathfrak{B}$.
- (ii) $\neg \mathfrak{A} \# \mathfrak{B} \leftrightarrow \mathfrak{A} = \mathfrak{B}$.
- (iii) $\mathfrak{A} \# \mathfrak{B} \rightarrow (\forall \mathfrak{C})(\mathfrak{A} \# \mathfrak{C} \vee \mathfrak{B} \# \mathfrak{C})$.

LEMMA 7.1. m intersects l & $D \in m \cap p$ & $D \omega l \rightarrow p$ intersects $l \vee p \neq m$

PROOF: There is a line m' through D so that $m' \parallel l$ (Th. 3); $m \neq m'$. $p \neq m$ or $p \neq m'$; if $p \neq m'$, then l intersects p (A3).

LEMMA 7.2: $C \in l$ & $C \omega m$ & $l \neq n$ & $l \cap m = l \cap n \rightarrow C \omega n$.

PROOF: Choose A so that $A \in n$ & $A \omega l$. $AC = p$. The line n intersects $l \vee n \neq p$ (Lemma 7.1). If n intersects l , then m intersects l , because $l \cap m = l \cap n$, so $C \omega n$ (A4). If $n \neq p$, then $C \omega n$ (A4).

PROOF OF THEOREM 7 (i): Choose l and m so that $\mathfrak{A} = \mathfrak{B}(l, m)$ and that $l \omega \mathfrak{B}$; choose C on l so that $C \omega m$. There is a line p so that $C \in p$, $p \in \mathfrak{B}$ (A1); $p \neq l$. Let n be a line in \mathfrak{A} ; $n \neq l \vee n \neq p$.

If $n \neq l$, then $C \omega n$ (Lemma 7.2), so $n \neq p$.

We have proved that $p \omega \mathfrak{A}$, so $\mathfrak{B} \neq \mathfrak{A}$.

PROOF OF THEOREM 7(ii): $\mathfrak{A} = \mathfrak{B}(l, m)$. Let p be a line in \mathfrak{B} , then $\neg p \omega \mathfrak{A}$. $p \neq l \vee p \neq m$; suppose $p \neq l$. I shall prove that $l \cap m = l \cap p$.

Suppose $X \in l \cap m$, so that \mathfrak{A} consists of all the lines through X . If we had $X \omega p$, then $p \omega \mathfrak{A}$, which gives a contradiction, so $X \in p$. Thus $l \cap m \subseteq l \cap p$.

Suppose $Y \in l \cap p$. I derive a contradiction from $Y \omega m$, as follows: Choose n in \mathfrak{A} ; $n \neq l \vee n \neq p$.

If $n \neq l$, then $Y \omega n$ (Lemma 7.2), so $n \neq p$.

Now $n \neq p$ for every n in \mathfrak{A} , so $p \omega \mathfrak{A}$. This is the desired contradiction. Thus $Y \in m$, and $l \cap p \subseteq l \cap m$.

PROOF OF THEOREM 7(iii): Choose l in \mathfrak{A} so that $l \omega \mathfrak{B}$; further m, r, s so that $\mathfrak{A} = \mathfrak{B}(l, m)$, $\mathfrak{C} = \mathfrak{B}(r, s)$. $l \neq r \vee l \neq s$; say $l \neq r$. As in the proof of part (i), we find a line p in \mathfrak{B} , so that $p \omega \mathfrak{A}$ and that p intersects l in D ; D can so be chosen that $D \omega r$ [If $D, E \in l$ and $D \neq E$, then $D \omega m$ or $E \omega m$; this is shown in the proof of Lemma 2.1. By choosing three points on l we find at least one point outside m and outside r]. Draw the line t so that $D \in t$, $t \in \mathfrak{C}$ (A1). $l \neq t \vee p \neq t$. Suppose $l \neq t$. For an arbitrary line u in \mathfrak{C} we have $u \neq l \vee u \neq t$. If $u \neq t$, then $D \omega u$ (Lemma 7.2), so $u \neq l$. It follows that $l \omega \mathfrak{C}$, so $\mathfrak{A} \neq \mathfrak{C}$. Similarly, if $p \neq t$, we have $\mathfrak{B} \neq \mathfrak{C}$.

THEOREM 8. $\mathfrak{A} \neq \mathfrak{B}$ & $l \in \mathfrak{A} \cap \mathfrak{B}$ & $m \in \mathfrak{A} \cap \mathfrak{B} \rightarrow l = m$.

PROOF: Suppose $l \neq m$; then $\mathfrak{A} = l \cap m$ and $\mathfrak{B} = l \cap m$. Apply Th. 2(ii).

7. Projective lines.

DEFINITION 9. $\mathfrak{A} \neq \mathfrak{B} \rightarrow \lambda(\mathfrak{A}, \mathfrak{B})$

$$= \{\mathfrak{C} | \mathfrak{A} \cap \mathfrak{C} = \mathfrak{A} \cap \mathfrak{B} \vee \mathfrak{B} \cap \mathfrak{C} = \mathfrak{A} \cap \mathfrak{B}\}.$$

If $\mathfrak{A} \neq \mathfrak{B}$, then $\lambda(\mathfrak{A}, \mathfrak{B})$ is a PROJECTIVE line (*p. line*); where $\lambda(\mathfrak{A}, \mathfrak{B})$ occurs, it is understood that $\mathfrak{A} \neq \mathfrak{B}$.

Greek lower case letters will be used to denote projective lines.

THEOREM 9. $\mathfrak{A} \neq \mathfrak{B} \ \& \ l \in \mathfrak{A} \cap \mathfrak{B} \rightarrow \lambda(\mathfrak{A}, \mathfrak{B}) = \{\mathfrak{C} | l \in \mathfrak{C}\}.$

PROOF: I. Let \mathfrak{C} be a p. point such that $l \in \mathfrak{C}$. $\mathfrak{C} \neq \mathfrak{A} \vee \mathfrak{C} \neq \mathfrak{B}$; suppose $\mathfrak{C} \neq \mathfrak{A}$. $l \in \mathfrak{A} \cap \mathfrak{B}$ and $l \in \mathfrak{A} \cap \mathfrak{C}$, so $\mathfrak{A} \cap \mathfrak{B} = \mathfrak{A} \cap \mathfrak{C}$ (Th. 8), so $\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B})$.

II. Let \mathfrak{P} be a p. point in $\lambda(\mathfrak{A}, \mathfrak{B})$. $\mathfrak{A} \cap \mathfrak{P} = \mathfrak{A} \cap \mathfrak{B}$ or $\mathfrak{B} \cap \mathfrak{P} = \mathfrak{A} \cap \mathfrak{B}$; in either case $l \in \mathfrak{P}$.

If, as in Th. 9, $\lambda(\mathfrak{A}, \mathfrak{B})$ contains a line l , then it is called a PROPER projective line; we write in this case $\lambda(\mathfrak{A}, \mathfrak{B}) = l$.

THEOREM 10. If $\mathfrak{A} \neq \mathfrak{B}$ and $\lambda(\mathfrak{A}, \mathfrak{B})$ is a proper p. line l , then either \mathfrak{A} or \mathfrak{B} is a proper p. point.

PROOF: Choose m so that $m \in \mathfrak{A}$, $m \omega \mathfrak{B}$, and C so that $C \in m$, $C \omega l$. $C \mathfrak{B} = n$. By A3, l intersects m or n , so \mathfrak{A} is proper or \mathfrak{B} is proper.

DEFINITION 10. A p. point \mathfrak{A} lies OUTSIDE the p. line $\lambda(\mathfrak{A} \omega \lambda)$, if \mathfrak{A} is apart from every p. point in λ .

THEOREM 11: $\mathfrak{A} \omega l$ is equivalent with $l \omega \mathfrak{A}$.

LEMMA 11.1: If $\mathfrak{A} \neq \mathfrak{B}$, \mathfrak{A} is proper $= A$, $A \mathfrak{B} = l$, $n \in \mathfrak{B}$, $n \neq l$, then $A \omega n$.

PROOF: Choose P so that $P \in n$, $P \omega l$. $AP = p$.

By Lemma 7.1, n intersects l or $n \neq p$.

If n intersects l , then \mathfrak{B} is proper, so $A \omega n$ by Axiom A4.

If $n \neq p$, then $A \omega n$, also by Axiom A4.

LEMMA 11.2: If $B \in l$ and $\mathfrak{A} \omega l$, then $B\mathfrak{A} \neq l$.

PROOF: $B\mathfrak{A} = p$; choose q in \mathfrak{A} so that $q \neq p$.

By Lemma 11.1, $B \omega q$, so $l \neq q$.

Put $\mathfrak{B}(l, q) = \mathfrak{C}$. By Th. 10, either \mathfrak{A} is proper ($\mathfrak{A} = A$) or \mathfrak{C} is proper ($\mathfrak{C} = C$).

If $\mathfrak{A} = A$, then $A \omega l$, so $AB \neq l$.

If $\mathfrak{C} = C$, then $C \omega p$ (Lemma 11.1), so $l \neq p$.

PROOF OF THEOREM 11: I. Suppose $\mathfrak{A} \omega l$. Choose B on l ; $B\mathfrak{A} = p$. By Lemma 11.2, $p \neq l$. Choose r in \mathfrak{A} ; $r \neq l$ or $r \neq p$.

If $r \neq p$, then $B \omega r$ (Lemma 11.1), so $l \neq r$.

We have proved that for every line r in \mathfrak{A} , $l \neq r$ is valid, so $l \omega \mathfrak{A}$.

II. Suppose $l \omega \mathfrak{A}$, and let \mathfrak{B} be any p. point of l ; then $l \in \mathfrak{B}$, so $\mathfrak{A} \neq \mathfrak{B}$. Thus $\mathfrak{A} \omega l$.

THEOREM 12: *If l is not apart from \mathfrak{A} , then $l \in \mathfrak{A}$.*

PROOF: This has been shown in the proof of Th. 7(ii).

DEFINITION 11. *Two p. lines λ and μ are APART from each other ($\lambda \neq \mu$) if there exists a p. point \mathfrak{A} so that $\mathfrak{A} \in \lambda$ and $\mathfrak{A} \omega \mu$.*

We shall not prove directly that the relation \neq between p. lines is an apartness relation; this follows from the main result of the paper, as it has been derived in [1, 2] from the axioms S1–S3, P1–P5.

8. Proof of the projective axioms. Our problem is to prove that projective points and projective lines satisfy the axioms P1–P5 of projective geometry. I have not succeeded in proving this from A1–A7; I must introduce some further axioms, which I shall mention where I want them.

P1. If the p. points \mathfrak{A} and \mathfrak{B} are apart from each other, then there is a p. line λ which contains \mathfrak{A} and \mathfrak{B} .

This is an immediate consequence of Def. 9.

P2. If the p. points \mathfrak{A} and \mathfrak{B} are apart from each other, and if both are contained in the p. lines λ and μ , then $\lambda = \mu$.

This will be proved in Theorem 23.

P3. If the p. lines λ and μ are apart from each other, then they have a p. point in common. This property, for the case that λ or μ is proper, is affirmed in a new axiom A8.

A8. $\mathfrak{A} \neq \mathfrak{B} \ \& \ l \omega \mathfrak{A} \rightarrow (\exists \mathfrak{C})(l \in \mathfrak{C} \ \& \ \mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B}))$.

The quantifier over projective points can be avoided. An equivalent

formulation is:

$$\begin{aligned} \text{A'8: } p \# q \ \& \ r \# s \ \& \ l \omega \mathfrak{P}(p, q) \ \& \ r \omega \mathfrak{P}(p, q) \rightarrow \\ (\exists t)[t \# l \ \& \ \mathfrak{P}(p, q) \cap \mathfrak{P}(r, s) = \mathfrak{P}(p, q) \cap \mathfrak{P}(t, l) \\ \vee \mathfrak{P}(p, q) \cap \mathfrak{P}(r, s) = \mathfrak{P}(r, s) \cap \mathfrak{P}(t, l)]. \end{aligned}$$

In the case that $\lambda(\mathfrak{A}, \mathfrak{B})$ is a proper p. line, A8 follows from A1–A7 and Def. 7. A8 suffices to prove P3, because it follows from Th. 17 below, that if the p. lines λ and μ are apart from each other, then λ or μ is a proper p. line.

We now turn to P4, the triangle axiom. First we consider the cases where two of the p. points $\mathfrak{A}, \mathfrak{B}, \mathfrak{C}$ are proper (Theorems 13, 14, 15).

THEOREM 13. *If $A \# B$ and $\mathfrak{C} \omega AB$, then $A \omega B\mathfrak{C}$.*

PROOF: $AB \omega \mathfrak{C}$ [Th. 11], so $AB \# B\mathfrak{C}$, so $A \omega B\mathfrak{C}$.

THEOREM 14. *If $A \# \mathfrak{B}$ and $C \omega A\mathfrak{B}$, then $A \omega C\mathfrak{B}$.*

PROOF: $A\mathfrak{B} = l$, $AC = p$, $C\mathfrak{B} = n$.

By Lemma 7.1, n intersects l or $n \# p$.

If n intersects l , then we can directly apply A4.

If $n \# p$, then $A \omega n$.

THEOREM 15. *If $A \# \mathfrak{B}$ and $C \omega A\mathfrak{B}$, then $\mathfrak{B} \omega AC$.*

PROOF: $AC = l$, $A\mathfrak{B} = m$, $C\mathfrak{B} = n$; $m \# n$.

Let p be any line in \mathfrak{B} . $p \# m$ or $p \# n$.

If $p \# m$, then $A \omega p$ (Lemma 11.1), so $p \# l$.

If $p \# n$, then $B \omega p$ (Lemma 11.1), so $p \# l$.

We have now proved that $l \omega \mathfrak{B}$, so $\mathfrak{B} \omega l$ (Theorem 11).

Let now at least one of the p. points be proper.

THEOREM 16. *If $A \# \mathfrak{B}$ and $\mathfrak{C} \omega A\mathfrak{B}$, then $\mathfrak{B} \omega A\mathfrak{C}$.*

PROOF: $A\mathfrak{B} = l$, $A\mathfrak{C} = p$.

$\mathfrak{C} \omega l$, so $l \omega \mathfrak{C}$ [Th. 11], so $l \# p$.

Let m be any line in \mathfrak{B} , $m \# l$ or $m \# p$.

If $m \# l$ we choose Q on m , so that $Q \omega l$, $AQ = q$.

m intersects $l \vee m \# q$. (Lemma 7.1).

If m intersects l , then \mathfrak{B} is proper, $\mathfrak{B} = B$, and $B \omega p$ by Th. 10, so $m \# p$.

If $m \# q$, then $A \omega m$, so $m \# p$.

We have proved that $m \# p$ for every line m in \mathfrak{B} , so $p \omega \mathfrak{B}$, so $\mathfrak{B} \omega p$ (Th. 11).

Note that Axiom A8 has not been used in the proofs of Theorems 13, 14, 15, 16.

There are two other cases of P4 in which only one of the three p. points is proper. I have not succeeded in proving these from the preceding axioms; therefore I introduce them as new axioms:

A9 *If $A \neq B$ and $C \omega AB$, then $A \omega \lambda(B, C)$.*

A10 *If $B \neq C$ and $A \omega \lambda(B, C)$, then $B \omega AC$.*

It follows from the next theorem that the case in which none of the three p. points is known to be proper, need not be considered.

THEOREM 17. *If $A \neq B$ and $C \omega \lambda(A, B)$, then at least one of the p. points A, B, C is proper.*

PROOF: Choose l so that $l \in C, l \omega A$.

By A8, there is a p. point D so that $l \in D, D \in \lambda(A, B)$. $D \neq C$; by Th. 10 C is proper or D is proper. If D is proper, $D = D$, then $\lambda(A, B) = D\mathcal{A}$ is proper, so, again by Th. 10, A is proper or B is proper.

It remains to prove the uniqueness of the p. line through two p. points which are apart from each other. We first prove some theorems about improper p. points.

THEOREM 18: *If A and B are improper p. points and $A \neq B$, then any line in A intersects any line in B .*

PROOF: Let l be a line in A and m a line in B . Choose p in A so that $p \omega B$, and choose C on p . $C\mathcal{B} = q$; then $q \neq p$. By Th. 16, $A \omega q$, so $q \omega A$, so $l \neq q$. p intersects q , so that we infer from A3 that l has a point in common with p or with q .

If l has a point in common with p , then l cannot be apart from p , because A is improper, so $l = p$, so l intersects q . It has now been proved that in every case l intersects q . Repeating this argument for B instead of A , we find that m intersects l .

THEOREM 19: *If $A \neq B, C \neq A$ and $C \cap B = A \cap B$, then $C \cap A = A \cap B$.*

PROOF: It is clear that $A \cap B \subseteq A \cap C$.

Let l be a line in $A \cap C$; we must prove that $l \in B$.

Case 1: A or B is a proper p. point. Then $\lambda(A, B)$ is a proper p. line m . $m \in A \cap B$, so $m \in A \cap C$. By Th. 8, $l = m$, so $l \in B$.

Case 2: (General case). Because $l \in A \cap C$, A or C is proper, so that

we may now assume that \mathfrak{C} is proper. Choose D on l so that $D \neq \mathfrak{B}$. $D\mathfrak{B} = n$. We shall give an indirect proof for $n = l$.

Suppose $n \neq l$; then it is impossible that \mathfrak{A} or \mathfrak{B} is a proper p. point, for in case 1 we know that $l \in \mathfrak{B}$, so $l = D\mathfrak{B} = n$. Thus \mathfrak{A} and \mathfrak{B} are both improper. It follows from Th. 10 that $\mathfrak{A} \cap \mathfrak{B} = \emptyset$, so $\mathfrak{B} \cap \mathfrak{C} = \emptyset$, so \mathfrak{C} is improper (A1). But \mathfrak{C} is proper; this contradiction proves that $n \neq l$ is impossible, so $n = l$, and $l \in \mathfrak{B}$.

COROLLARY. *In the case that $\mathfrak{C} \neq \mathfrak{A}$ we have*

$$\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B}) \leftrightarrow \mathfrak{A} \cap \mathfrak{C} = \mathfrak{A} \cap \mathfrak{B}.$$

THEOREM 20. *If \mathfrak{A} and \mathfrak{B} are improper p. points, and $\mathfrak{A} \neq \mathfrak{B}$, then $\lambda(\mathfrak{A}, \mathfrak{B})$ is the set of all improper p. points.*

PROOF: By Th. 10, $\mathfrak{A} \cap \mathfrak{B} = \emptyset$. Let \mathfrak{P} be any p. point in $\lambda(\mathfrak{A}, \mathfrak{B})$; as $\mathfrak{A} \neq \mathfrak{P}$ or $\mathfrak{B} \neq \mathfrak{P}$, we may assume that $\mathfrak{A} \neq \mathfrak{P}$.

Then, by Th. 19, $\mathfrak{A} \cap \mathfrak{P} = \mathfrak{A} \cap \mathfrak{B} = \emptyset$, so \mathfrak{P} is improper. Conversely, if \mathfrak{P} is improper, we have by Th. 10, that $\mathfrak{A} \cap \mathfrak{P} = \emptyset = \mathfrak{A} \cap \mathfrak{B}$, so $\mathfrak{P} \in \lambda(\mathfrak{A}, \mathfrak{B})$.

The following theorem asserts the uniqueness of the p. point of which the existence is affirmed in A8.

THEOREM 21. *If $\mathfrak{A} \neq \mathfrak{B}$, \mathfrak{C} and \mathfrak{D} belong to $\lambda(\mathfrak{A}, \mathfrak{B})$, $p \omega \mathfrak{A}$ and p belongs to \mathfrak{C} and to \mathfrak{D} , then $\mathfrak{C} = \mathfrak{D}$.*

PROOF: Suppose $\mathfrak{C} \neq \mathfrak{D}$. As $\mathfrak{C} \neq \mathfrak{A}$ and $\mathfrak{D} \neq \mathfrak{A}$, it follows from Th. 19 that $\mathfrak{A} \cap \mathfrak{B} = \mathfrak{A} \cap \mathfrak{C} = \mathfrak{A} \cap \mathfrak{D}$. Thus, if $\mathfrak{A} \cap \mathfrak{B}$ contained a line l , we should have $l \in \mathfrak{C} \cap \mathfrak{D}$, so, by Th. 8, $l = p$; but this contradicts $p \omega \mathfrak{A}$. It follows that $\mathfrak{A} \cap \mathfrak{B} = \emptyset$.

Thus \mathfrak{A} , \mathfrak{B} , \mathfrak{C} , \mathfrak{D} are all improper p. points (Th. 20), and, as $\mathfrak{C} \neq \mathfrak{D}$, $\mathfrak{C} \cap \mathfrak{D} = \emptyset$, contradicting $p \in \mathfrak{C} \cap \mathfrak{D}$. We have proved that $\mathfrak{C} \neq \mathfrak{D}$ is impossible, so $\mathfrak{C} = \mathfrak{D}$.

THEOREM 22. *If $\mathfrak{A} \neq \mathfrak{B}$ and if it is impossible that $\mathfrak{C} \omega \lambda(\mathfrak{A}, \mathfrak{B})$, then $\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B})$.*

PROOF: We treat the case that $\mathfrak{C} \neq \mathfrak{A}$. Choose p in \mathfrak{C} so that $p \omega \mathfrak{A}$. By A8, there is a p. point \mathfrak{G} so that $p \in \mathfrak{G}$, $\mathfrak{G} \in \lambda(\mathfrak{A}, \mathfrak{B})$. We give an indirect proof for $\mathfrak{C} = \mathfrak{G}$. Suppose that $\mathfrak{C} \neq \mathfrak{G}$. As $p \in \mathfrak{C} \cap \mathfrak{G}$, \mathfrak{C} or \mathfrak{G} is a proper p. point (Th. 10). If \mathfrak{G} is proper, $\mathfrak{G} = G$, then $\lambda(\mathfrak{A}, \mathfrak{B})$ is proper, $\lambda(\mathfrak{A}, \mathfrak{B}) = l$; $G\mathfrak{A} = l$, $G\mathfrak{C} = p$. $\mathfrak{A} \omega \mathfrak{G}\mathfrak{C}$, so $\mathfrak{C} \omega G\mathfrak{A}$, $\mathfrak{C} \omega \lambda(\mathfrak{A}, \mathfrak{B})$, which is impossible by hypothesis. If \mathfrak{C} is proper, $\mathfrak{C} = C$, then $\lambda(\mathfrak{A}, \mathfrak{C})$ is proper,

$\lambda(\mathfrak{A}, \mathfrak{C}) = m$. Suppose that $\mathfrak{B} \omega m$; then we should have $C \omega \lambda(\mathfrak{A}, \mathfrak{B})$ (A9), which is impossible by hypothesis. But if $\mathfrak{B} \in m$, then $\lambda(\mathfrak{A}, \mathfrak{B}) = m$, $\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B})$, $\mathfrak{C} = \mathfrak{G}$. We have derived a contradiction from the hypothesis that $\mathfrak{C} \neq \mathfrak{G}$, so $\mathfrak{C} = \mathfrak{G}$, and $\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B})$.

Remark: A9 is only used in this proof, A10 nowhere in this paper. However, A10 will be needed for the derivation of further theorems of projective geometry.

The next theorem asserts P2 for p. points.

THEOREM 23. *If $\mathfrak{A} \neq \mathfrak{B}$, $\mathfrak{A} \neq \mathfrak{C}$ and $\mathfrak{C} \in \lambda(\mathfrak{A}, \mathfrak{B})$, then $\lambda(\mathfrak{A}, \mathfrak{C}) = \lambda(\mathfrak{A}, \mathfrak{B})$.*

PROOF: We first prove the theorem, making the extra assumption that $\lambda(\mathfrak{A}, \mathfrak{B})$ is a proper p. line l . In this case, $l \in \mathfrak{A} \cap \mathfrak{B}$, so $\lambda(\mathfrak{A}, \mathfrak{B}) = \{\mathfrak{P} | l \in \mathfrak{P}\}$. Moreover, $l \in \mathfrak{A} \cap \mathfrak{C}$, so $\lambda(\mathfrak{A}, \mathfrak{C}) = \{\mathfrak{P} | l \in \mathfrak{P}\}$. Thus $\lambda(\mathfrak{A}, \mathfrak{B}) = \lambda(\mathfrak{A}, \mathfrak{C})$.

In the general case, let \mathfrak{D} be any p. point in $\lambda(\mathfrak{A}, \mathfrak{B})$. Suppose that $\mathfrak{D} \omega \lambda(\mathfrak{A}, \mathfrak{C})$; then, by Th. 17, at least one of the p. points \mathfrak{A} , \mathfrak{C} , \mathfrak{D} is proper, so that $\lambda(\mathfrak{A}, \mathfrak{B})$ is proper, but in this case we have already proved that $\lambda(\mathfrak{A}, \mathfrak{B}) = \lambda(\mathfrak{A}, \mathfrak{C})$. Thus the assumption that $\mathfrak{D} \omega \lambda(\mathfrak{A}, \mathfrak{C})$ has led to a contradiction; by Th. 22, $\mathfrak{D} \in \lambda(\mathfrak{A}, \mathfrak{C})$. We have now proved that $\lambda(\mathfrak{A}, \mathfrak{B}) \subseteq \lambda(\mathfrak{A}, \mathfrak{C})$. In particular, $\mathfrak{B} \in \lambda(\mathfrak{A}, \mathfrak{C})$. Now we prove by the same argument, interchanging \mathfrak{B} and \mathfrak{C} , that $\lambda(\mathfrak{A}, \mathfrak{C}) \subseteq \lambda(\mathfrak{A}, \mathfrak{B})$.

Bibliography

- [1] HEYTING, A., *Intuitionistische axiomatiek der projectieve meetkunde*. Thesis University of Amsterdam. Groningen 1925.
- [2] —, *Zur intuitionistischen Axiomatik der projektiven Geometrie*. Mathematische Annalen, vol. 98 (1927), pp. 491–538.
- [3] —, *Die formalen Regeln der intuitionistischen Logik*, Sitzungsberichte preuss. Akad. Wiss. Berlin (1930), pp. 42–56.
- [4] —, *Die formalen Regeln der intuitionistischen Mathematik*, Sitzungsberichte preuss. Akad. Wiss. Berlin (1930), pp. 55–71.
- [5] —, *Intuitionism, an introduction*. Amsterdam 1956.

GRUNDLAGEN DER GEOMETRIE VOM STANDPUNKTE DER ALLGEMEINEN TOPOLOGIE AUS

KAROL BORSUK

Universität Warschau, Warschau, Polen

Das Problem die Geometrie axiomatisch zu begründen, das zum ersten Mal von Euklid gestellt und gelöst wurde, hat bis jetzt seine Aktualität nicht verloren. Am Ende des 19. Jahrhunderts hat Hilbert [9] die berühmte Axiomatik der Geometrie angegeben, womit er eine wesentliche Vertiefung und Vervollständigung der Ideen von Euklid erzielt und der Geometrie die Gestalt einer deduktiven Theorie, im modernen Sinne, gegeben hat. Durch die Axiomatik von Hilbert ist auch das Verhältnis zwischen den Euklidischen und der hyperbolischen Geometrie von Łobatschewsky und Bolyai endgültig erklärt.

Dadurch wurde aber das Problem der Grundlagen der Geometrie keineswegs auserschöpft. Einerseits, öffnete die Entwicklung der mathematischen Logik weite Möglichkeiten für die Untersuchung der logischen Struktur der Geometrie, als einer deduktiven Theorie. In dieser Richtung geht ein bedeutender Teil der modernen Studien auf dem Gebiete der Grundlagen der Geometrie. Andererseits, durch die Einführung von verschiedenen Typen der allgemeinen Räume ist das Problem entstanden, die Lage der klassischen Räume unter sämtlichen abstrakten Räumen aufzuklären.

Das Problem die klassischen Geometrien unter sämtlichen Riemannschen Geometrien zu charakterisieren, wurde schon im 19. Jahrhundert von Sophus Lie [15] gestellt. Lie, zu den früheren Ideen von Helmholtz [8] anknüpfend, hat diesem Problem die Gestalt des Problems einer Charakterisierung der Gruppe der starren Bewegungen gegeben. Aber erst die Entstehung der allgemeinen Mengenlehre und der auf ihr gestützten axiomatischen Theorie der allgemeinen, abstrakten Räume, hat dem Problem der Grundlagen der Geometrie eine wirklich allgemeine, moderne Gestalt gegeben. Von diesem allgemeinen Standpunkte aus wurde das von Lie gestellte Problem die Geometrie mit Hilfe der Gruppe der starren Bewegungen zu begründen, im Jahre 1930 von Kolmogoroff [11] angegriffen. Kolmogoroff hat ein System von Axiomen angegeben, durch

das die Klasse der Riemannschen Räume mit der konstanten Krümmung charakterisiert wurde. Leider hat Kolmogoroff die vollständigen Beweise seiner Behauptungen nicht veröffentlicht und erst in den letzten Jahren sind zwei Arbeiten von Tits [20], [21] und dann eine Arbeit von Freudenthal [7] erschienen, in den eine, im gewissen Sinne endgültige Lösung des Raumproblems auf dem Boden der Charakterisierung der Bewegungsgruppe gegeben ist. In der Arbeit von Freudenthal ist auch eine weitgehende Klassifikation der auf diese Weise charakterisierten Geometrien angegeben.

In eine etwas andere Richtung gehen die Arbeiten, die zu einer Charakterisierung der klassischen Räume auf dem Boden der durch Menger [17] entwickelten allgemeinen metrischen Geometrie streben. Ausser den wesentlichen Ergebnissen von Menger [17], [18], soll man hier die Ergebnisse von Wilson [26], von Garret Birkhoff [2], von Blumenthal [3], von H. C. Wang [25] und von anderen nennen. Wilson charakterisierte die Euklidischen Metriken mit Hilfe einer gewissen metrischen Eigenschaft jeder vier Punkten des Raumes. Garret Birkhoff, und die anderen, stützen ihre Untersuchungen auf dem Postulate einer metrischen Homogenität.

In diesem Vortrage möchte ich mich mit einer Charakterisierung der klassischen Räume auf dem Boden einer Klassifikation der allgemeinen topologischen Räume beschäftigen. Es handelt sich dabei vor allem um eine Formulierung des Problems und um die Andeutung der hier entstehenden Schwierigkeiten. Ich bin nicht imstande eine definitive Lösung des Problems anzugeben und ich meine sogar, dass wir noch fern davon sind. Ich möchte nur eine partielle Lösung angeben, die als eine Illustration der allgemeinen Tendenz dieser Betrachtungen dienen kann.

Um die Grundlagen der Geometrie auf einer breiten Basis der topologischen Eigenschaften zu bauen, braucht man eine mehr rausgebaute Systematik der topologischen Räume zu bearbeiten. Bis jetzt ist eine solche Systematik wenig entwickelt. Am besten ist es mit der Systematik der allgemeinsten Typen der topologischen Räume. Verschiedene Axiome: der Regelmässigkeit, der Normalität, der Basis und so weiter erlauben gewisse Klassen von Räumen mit mehr oder weniger reichem geometrischen Inhalt zu definieren. Aber die Klassifikation von den mehr speziellen Räumen ist bis jetzt höchst mangelhaft und wir sind immer fern von der topologischen Bestimmung der wichtigen Klasse der sogenannten Polyeder. Ich verstehe dabei hier, unter den *Polyedern*, solche separable Räume, für die eine lokal endliche Triangulation existiert. Ich glaube, dass erst eine Entwicklung der Systematik von den topologischen Räumen,

eine angemessene Grundlage zur genauen Aufklärung der Natur der klassischen Räume schaffen wird. Da aber die Geometrie, neben topologischen, auch metrische Axiome braucht, so scheint mir, dass eine für oben genannte Zwecke nützliche Axiomatik, neben topologischen auch metrische Axiome enthalten soll und zwar unter der Berücksichtigung folgendes „*Prinzips eines topologisch-metrischen Parallelismus*“:

PRINZIP $[T||M]$. *Die topologischen Axiome sollen die Existenz einer den metrischen Axiomen genügenden Metrik implizieren. Die metrischen Axiome sollen die Erfüllung der topologischen Axiome implizieren.*

Die wohlbekannten Schwierigkeiten bei Versuchen die Euklidischen Räume topologisch zu charakterisieren haben zur Folge, dass eine vollständige Realisierung des Prinzips $[T||M]$, bei aktuellem Niveau der Topologie, eher aussichtslos ist. Man kann aber den Entwurf einer Axiomatik der Euklidischen Geometrie angeben, einer Axiomatik, die wenigstens teilweise dieses Prinzip berücksichtigt.

Es ist zweckmässig unsere Axiome in drei folgende Gruppen zu teilen:

- I. *Axiome der allgemeinen Räume.*
- II. *Axiome der Regelmässigkeit.*
- III. *Spezielle Axiome.*

Als die zur ersten Gruppe gehörenden Axiome kann man irgendeine Axiome, die die metrisierbaren, separablen Räume charakterisieren, wählen. Man kann, zum Beispiel, die drei *Axiome der abgeschlossenen Hülle* (von Kuratowski [13], [14]), das *Axiom der Normalität* und das *Axiom der abzählbaren Basis* nehmen. Wenn wir, wie üblich, die abgeschlossene Hülle der Menge X mit \bar{X} bezeichnen, so hat die erste Gruppe der topologischen Axiome die folgende Gestalt (siehe [14]):

AXIOME (I, T) .

- (1) $\overline{X \cup Y} = \bar{X} \cup \bar{Y}$.
- (2) Falls X leer oder einpunktig ist, so ist $\bar{X} = X$.
- (3) $\bar{\bar{X}} = \bar{X}$.
- (4) Für je zwei disjunkte, abgeschlossene Mengen X und Y gibt es eine offene Menge G von der Art, dass $X \subset G$ und $\bar{G} \cap Y = \emptyset$ ist (Axiom der NORMALITÄT).
- (5) Es gibt eine Folge $\{G_n\}$ von offenen Mengen von der Art, dass jede offene Menge Vereinigungsmenge gewisser Mengen dieser Folge ist (Axiom der ABZÄHLBAREN BASIS).

Als die entsprechende Gruppe der metrischen Axiome nehmen wir die drei *Axiome der metrischen Räume* von Fréchet (mit einer Modifikation von Lindenbaum [16]) und das *Axiom der Separabilität*. Wenn wir, wie üblich, die Entfernung von dem Punkte x bis dem Punkte y mit $\rho(x, y)$ bezeichnen, so hat die erste Gruppe der metrischen Axiome die folgende Gestalt:

AXIOME (I, M) .

- (1) $\rho(x, y)$ ist reell.
- (2) $\rho(x, y) = 0$ dann und nur dann wenn $x = y$.
- (3) $\rho(x, y) + \rho(x, z) \geq \rho(y, z)$.
- (4) Es gibt eine im Raume dichte, höchstens abzählbare Teilmenge (Axiom der SEPARABILITÄT).

Der wohlbekannte Metrisationssatz von Urysohn [23] besagt, dass die Axiome (I, T) die Existenz einer den Axiomen (I, M) genügenden Metrik implizieren. Auch umgekehrt, die Existenz einer Metrik, die den Axiomen (I, M) genügt, hat zur Folge die Erfüllung sämtlicher Axiome der Gruppe (I, T) . Somit bei diesen Axiomen ist das Prinzip des topologischen metrischen Parallelismus erfüllt.

Die Räume, die der ersten Gruppe der Axiome genügen, bilden eine wichtige und gut bekannte Klasse. Aber diese Klasse ist so allgemein, dass sie auch viele Räume mit recht komplizierten und wenig anschaulichen Eigenschaften enthält. Die zweite Gruppe von Axiomen soll unter den allgemeinen metrischen Räumen eine Klasse von Räumen mit besonders einfachen, anschaulichen Eigenschaften bestimmen. Diese Axiome sollen somit verschiedene, sogenannte *pathologische Phänomene* eliminieren [5]. Zusammen mit den Axiomen der ersten Gruppe, sollen sie eine topologische Grundlage für jede „vernünftige“ Geometrie bilden. Im Gegensatz zu der genau bestimmten ersten Gruppe von Axiomen, die Aufstellung der Axiome der zweiten Gruppe ist wenig bestimmt. Es scheint, dass diese Axiome vor allem die lokalen Eigenschaften des Raumes anbetreffen sollen und eine Klasse der Räume definieren, die hinreichend umfangreich sein soll um alle Polyeder zu enthalten, aber hinreichend speziell, um alle Räume mit paradoxalen Eigenschaften ausschliessen. Ich werde hier diese Gruppe von Axiomen nur provisorisch folgendermassen aufstellen:

AXIOME (II, T) .

- (1) *Lokale Kompaktheit.*
- (2) *Lokaler Zusammenhang.*

Sicher ist die so aufgestellte Axiomgruppe nicht hinreichend um die „vernünftige Räume“ zu charakterisieren. Für unseren bescheidenen Zweck, einen sehr unvollkommenen Prototypus einer topologisch-metrischen Axiomatik zu geben, wird sie aber genügen.

Die entsprechende Gruppe der metrischen Axiome besteht aus zwei Axiome:

AXIOME (II, M).

- (1) *Kompaktheit der beschränkten abgeschlossenen Mengen.*
- (2) *Lokale Konvexität.*

Man sagt dabei, dass ein Raum X *lokal konvex* ist, wenn für jeden Punkt $a \in X$ eine Umgebung U existiert von der Art, dass für je zwei Punkte $x, y \in U$ mindestens einen Punkt $z \in X$ gibt, für den

$$\rho(x, z) = \rho(y, z) = \frac{1}{2} \cdot \rho(x, y)$$

gilt. Jeder solche Punkt z soll ein *Mittelpunkt* des Paares x, y heissen.

Es ist bekannt (vgl. Menger) dass in einem metrischen Raume X , in dem die Axiome (II, M) erfüllt sind, gibt es für jeden Punkt $a \in X$ eine positive Zahl r von der Art, dass jeder Punkt $x \in X - (a)$, der eine Entfernung von a kleiner als r hat, mit a durch eine geradlinige Strecke verbunden sein kann. Daraus ergibt sich ohne Weiteres, dass die Axiome (II, M) die Axiome (II, T) zur Folge haben. Ob auch umgekehrt, die Axiome (II, T) (zusammen mit den Axiomen der ersten Gruppe) die Existenz einer der Axiomen (II, M) genügenden Metrik zur Folge haben, ist noch nicht endgültig aufgeklärt. Die Ergebnisse von Bing [1], der die von Menger ausgesprochene Vermutung der konvexen Metrisierbarkeit der lokal zusammenhängenden Kontinua bewiesen hat, und auch die neulich erhaltenen Ergebnisse von japanischen Mathematikern Tanaka und Tominaga [19], [22], erlauben aber zu vermuten, dass bei diesen Axiomen das Prinzip $[T||M]$ erfüllt wird. Da aber die zweite Gruppe der Axiome sicher nicht als definitiv aufgestellt gelten kann, so meine ich dass die wesentlichen Schwierigkeiten erst bei einer angemessenen Vervollständigung dieser Gruppe der Axiome erscheinen werden.

Die Axiome der dritten Gruppe sollen die spezifischen Eigenschaften des anbetreffenden Raumes angeben. Falls wir das Prinzip des topologisch-metrischen Parallelismus realisieren wollen, so sollen die Axiome (III, T), zusammen mit den Axiomen (I, T) und (II, T), den betrachteten Raum topologisch vollständig charakterisieren. Eine topologische Charakterisierung der Euklidischen Ebene war schon vor vielen Jahren von

van Kampen [10] gegeben. Somit, in diesem Spezialfalle, bietet eine Aufstellung der dem Prinzip $[T|M]$ genügenden Axiomatik keine Schwierigkeiten. Eine ähnliche Axiomatik für höherdimensionale Euklidische Räume zu finden ist eine unvergleichlich schwierigere Aufgabe. Es zeigt sich aber möglich, die Axiomgruppen (III, T) und (III, M) so anzugeben, dass sie mitgesamt und mit den Axiomen der ersten und zweiten Gruppe eine vollständige Axiomatik der elementaren Geometrie bilden. Das Prinzip $[T|M]$ wird dabei vernachlässigt, weil die von ihm verlangte Aussonderung der topologischen und metrischen Axiome nicht erfüllt wird. Die Schwierigkeit eine solche Aussonderung zu realisieren liegt hauptsächlich auf der topologischen Seite. Somit wird man desto näher zu der Realisierung des Prinzipes $[T|M]$ kommen, je reicher der Inhalt der topologischer Axiome (III, T) wird. Anders gesprochen, ist es zweckmässig die Rolle der metrischen Axiome (III, M) möglichst weit zu reduzieren.

Um die Axiome der Gruppe (III, T) für Euklidische Räume zu formulieren, werde ich folgenden Begriff benutzen:

Ein *wahrer Zyklus* (im Sinne von Vietoris [24]) γ ist im Raume X mit dem Punkte $x \in X$ verschlungen, wenn γ einen kompakten Träger $A \subset X - (x)$ hat und γ nicht homolog Null in der Menge $X - (x)$ ist.

Nun besteht die Axiomgruppe (III, T) von vier folgenden Axiomen:

AXIOME (III, T) .

- (1) *Der Raum ist zusammenhängend.*
- (2) *Die Dimension des Raumes ist gleich n .*
- (3) *In jeder Umgebung jedes Punktes des Raumes gibt es einen wahren Zyklus, der in dem Raume mit diesem Punkte verschlungen ist.*
- (4) *Für jeden wahren Zyklus des Raumes ist die Menge der Punkte mit den dieser Zyklus verschlungen ist, offen.*

Offenbar sind die Axiome der Gruppen (I, T) , (II, T) , (III, T) nicht hinreichend um den Euklidischen n -dimensionalen Raum topologisch zu charakterisieren. Sie sind erfüllt, zum Beispiel, durch jede offene n -dimensionale Mannigfaltigkeit und auch durch verschiedene andere Räume. Man kann sie in verschiedene Weisen verstärken. Da wir aber, wie ich schon gesagt habe, eine rein topologische Charakterisierung des n -dimensionalen Euklidischen Raumes nicht angeben können, sind wir gezwungen das Prinzip $[T|M]$ vernachlässigend, gewisse metrische Bedingungen einzuführen. Um den n -dimensionalen Euklidischen Raum voll-

ständig zu charakterisieren, genügt es zu den oben genannten Axiomen folgendes metrische Axiom (III, M) hinzufügen:

AXIOM (III, M).

Jede vier Punkte des Raumes sind zur gewissen vier Punkten des Euklidischen 3-dimensionalen Raumes E_3 kongruent.

Dieses Axiom, das von Menger [17] formuliert und von Wilson [26] und den anderen benutzt war wird, wie üblich, *Vierpunktebedingung* genannt.

Die aus den drei Gruppen der topologischen und der metrischen Axiome bestehende Axiomatik des Euklidischen n -dimensionalen Raumes ist in Wirklichkeit nur eine Modifikation der rein metrischen Axiomatik, die im Jahre 1932 von Wilson [26] angegeben war. Der Zweck dieser Modifikation war, die metrischen Axiome in möglichst grossen Masse durch die topologischen zu ersetzen um somit zur Realisierung des Prinzips $[T||M]$ näher zu kommen.

Nun werden wir zeigen, dass unsere Axiomatik den Euklidischen n -dimensionalen Raum E_n vollständig charakterisiert.

Aus den Axiomen der ersten und zweiten Gruppe folgt, dass für jeden Punkt $a \in X$ eine offene Umgebung U existiert von der Art, dass jeder Punkt $b \in U$ mit a durch eine geradlinige Strecke sich vereinigen lässt. Wir können dabei annehmen, dass diese Umgebung beschränkt ist. Aus der Vierpunktbedingung ergibt sich, dass die a und b verbindende Strecke eine einzige ist. Wir werden sie mit \overline{ab} bezeichnen. Da U beschränkt ist, ist auch die Vereinigung aller dieser Strecken beschränkt. Wir schliessen leicht, indem wir das erste von den Axiomen (II, M) anwenden, dass die Strecke \overline{ab} stetig von ihren Endpunkten a und b abhängt.

Um nun zu zeigen, dass jede Strecke \overline{ab} sich aussen den Endpunkt b verlängern lässt, betrachten wir eine Umgebung V des Punktes b so klein, dass sie den Punkt a nicht enthält und dass je zwei ihre Punkte sich durch eine eindeutig bestimmte und von ihren Endpunkten stetig abhängende Strecke vereinigen lassen. Nach dem dritten der Axiome (III, T) gibt es in der Umgebung V einen wahren Zyklus γ der mit b verschlungen ist (vgl. Abb. 1). Betrachten wir einen kompakten Träger $B \subset V - (b)$ und einen Punkt $c \in \overline{ab} \cap V$, der von b verschieden ist. Man sieht leicht, dass der wahre Zyklus γ in der Vereinigungsmenge aller Strecken \overline{cx} mit $x \in B$, homolog Null ist. Da γ nicht homolog Null in der Menge $X - (b)$ ist schliessen wir, dass es einen Punkt $b' \in B$ gibt von der Art, dass die Strecke $\overline{cb'}$ den Punkt b enthält. Mit Hilfe der Vierpunktebedingung

schliessen wir, dass $\overline{ac} \cup \overline{cb'}$ die gesuchte Verlängerung der Strecke \overline{ab} ist. Daraus ergibt sich leicht (durch Anwendung des ersten der Axiome (II, M)) dass es eine Halbgerade gibt, die a als ihren Endpunkt hat und den Punkt b enthält. Mit Hilfe des ersten der Axiome (II, M) und der Vierpunktebedingung zeigt man ferner, dass diese Halbgerade stetig vom Punkte b abhängt in dem Sinne, dass der um $s > 0$ von a entfernte Punkt dieser Halbgerade stetig von s und b abhängt.

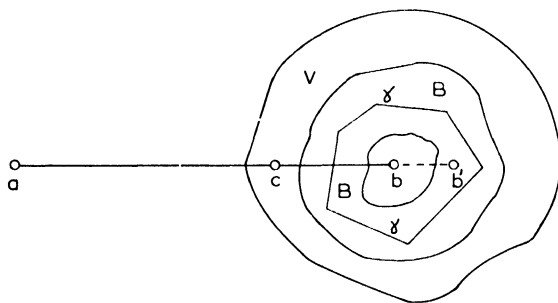


Fig. 1

Nun werden wir zeigen, dass es für jeden Punkt $x \in X$ eine Halbgerade H gibt, die a als ihren Endpunkt hat und den Punkt x enthält. Nach dem zweiten der Axiome (II, M) gibt es eine Zahl $r > 0$ von der Art, dass eine solche Halbgerade für alle Punkte x mit $\rho(a, x) \leq r$ existiert. Es bezeichne Q die Vollkugel um a mit dem Radius r . Wir setzen voraus, dass

$$\rho(a, x) > r > 0$$

gilt. Da der Raum X zusammenhängend (Axiom (III, T), 1), lokal zusammenhängend (Axiom (II, T) 2) und lokal kompakt (Axiom (II, T), 1) ist, gibt es in X einen einfachen Bogen B , der a und x vereinigt. Es sei s eine Zahl, die grösser als der Durchmesser von B ist. Nach dem dritten der Axiome (III, T) gibt es in Q einen wahren Zyklus γ , der mit a verschlungen ist. Es bezeichne A den in $Q - (a)$ enthaltenen, kompakten Träger von γ (vgl. Abb. 2). Da Q in sich selbst zu dem Punkte a zusammenziehbar ist, ist dieser Zyklus in Q homolog Null. Somit ist er mit dem Punkte x nicht verschlungen.

- (i) γ ist verschlungen mit a ,
- (ii) γ ist nicht verschlungen mit x .

Nun nehmen wir an, dass keine der Halbgeraden mit dem Endpunkt a

den Punkt x enthält. Für jeden Eckpunkt p des Zyklus γ bezeichnen wir mit $\varphi(p)$ den um s von a entfernten Punkt der Halbgerade \overrightarrow{ap} . Da die betrachtete Halbgerade stetig vom Punkte p abhängt schliessen wir, dass φ den wahren Zyklus γ auf einen wahren Zyklus γ^* abbildet. Dabei sind die wahren Zyklen γ und γ^* in der Vereinigungsmenge aller Strecken $\overline{p\varphi(p)}$, wo $p \in A$, homolog. Da aber diese Vereinigungsmenge keinen der

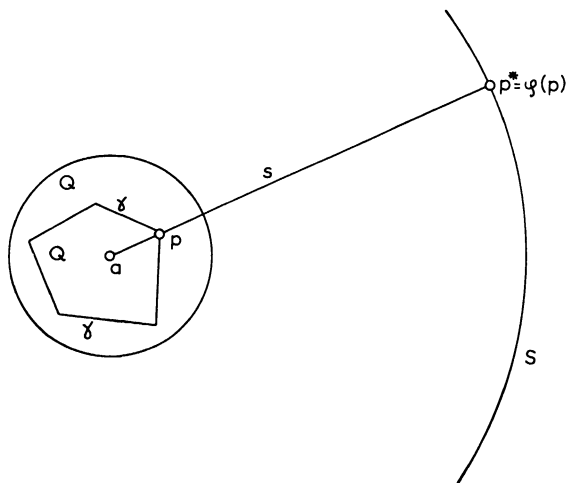


Fig. 2

Punkte a und x enthält, schliessen wir, dass

(iii) $\gamma \sim \gamma^*$ in $X - (a) - (x)$

gilt. Daraus und aus (i) und (ii), folgt

(iv) γ^* ist verschlungen mit a ,

(v) γ^* ist nicht verschlungen mit x .

Aber der Zyklus γ^* liegt auf der Oberfläche S der Kugel um den Punkt a mit dem Radius s . Da s von dem Diameter $\delta(B)$ von B grösser ist folgt, dass B mit S punktfremd ist. Wenn wir nun (iv) und das vierte der Axiome (III, T) beachten, so sehen wir dass der Zyklus γ^* mit dem Punkte x verschlungen sein soll, was der Bedingung (v) widerspricht.

Somit haben wir gezeigt, dass je zwei verschiedene Punkte unseres Raumes auf einer Geraden liegen. Wie aber W. A. Wilson gezeigt hatte [26] ist ein metrischer, n -dimensionaler, separabler Raum, in dem je zwei

Punkte auf einer Geraden liegen, wobei das Axiom (III, M) erfüllt ist, mit dem n -dimensionalen Euklidischen Raume E_n kongruent. Somit sehen wir, dass unsere Axiome den Raum E_n vollständig charakterisieren.

Der Hauptmangel der angegebenen Axiomatik besteht darin, dass in der dritten Gruppe der Axiome die topologischen und die metrischen Voraussetzungen nicht getrennt sind. Ich habe schon erwähnt, dass bei dem aktuellen Stande der Topologie diese Trennung als aussichtslos betrachtet werden kann.

Im Falle der Ebene stellt sich aber die Sache anders. In diesem Falle können wir die von van Kampen angegebene topologische Charakterisierung der Euklidischen Ebene verwenden [10]. Ich werde diese Axiomatik in einer etwas modifizierten Gestalt angeben, um die Verwendung der in der ursprünglichen Axiomatik von van Kampen gebrauchten speziellen Begriffen eines einfachen Bogens und einer einfachen geschlossenen Kurve zu vermeiden. In dieser modifizierten Gestalt, besteht die topologische Axiomatik der Ebene aus den Axiomen (I, T), (II, T) und aus der folgenden Gruppe der speziellen Axiome:

AXIOME (III, T)₂.

- (1) *Der Raum ist zusammenhängend.*
- (2) *Der Raum ist nicht kompakt.*
- (3) *Jeder kompakter Schnitt des Raumes ist nicht azyklisch.*
- (4) *Jedes Teilkompaktum des Raumes, das nicht azyklisch in der Dimension 1 ist, ist ein Schnitt.*

Um das Prinzip $[T||M]$ zu realisieren, genügt es nun als metrische Axiome die folgenden Axiome nehmen:

AXIOME (III, M)₂.

- (1) *Je zwei Punkte liegen auf einer Geraden.*
- (2) *Vierpunktebedingung.*

Zum Schluss dieses Vortrages möchte ich einige Bemerkungen allgemeines Natur hinzufügen. Das Problem der Grundlagen der Geometrie habe ich hier als ein Fragment des allgemeinen *Problems der Klassifikation* der topologischen Räume aufgefasst. Von diesem Standpunkte aus soll man auch die hier angegebenen drei Axiomgruppen betrachten. Ich habe schon bemerkt, dass die zweite Gruppe, die ich die *Gruppe der Regelmässigkeit* genannt habe, hier nur provisorisch aufgestellt war. Sie soll unter sämtlichen topologischen Räumen eine Klasse von Räumen mit be-

sonders regelmässigen Eigenschaften bestimmen. Eine volle topologische Charakterisierung der Klasse von Polyedern bietet sehr wesentliche Schwierigkeiten — da dadurch eine Überbrückung des Abgrundes zwischen den axiomatisch definierten abstrakten topologischen Räumen und den durch Konstruktion definierten Figuren der elementaren Geometrie erzielt würde. Nur im Falle der höchstens zweidimensionalen Polyedern ist es neulich gelungen diese Schwierigkeiten zu überwinden (Kosiński [12]). Dagegen ist auch in dem allgemeinen Falle eine axiomatische Auffassung eines gewissen Teiles der topologischen Eigenschaften der Polyeder sicher möglich (vgl. [4]). Es entsteht aber die Frage, wie man diesen „gewissen Teil“ definieren soll. Diese Frage ist eng mit der Frage der vernünftigen Klassifikation der topologischen Invarianten verbunden.

Seit dem *Erlanger-Programm* von Felix Klein, klassifiziert man verschiedene geometrische Eigenschaften vom Standpunkte der Klassen der Abbildungen, gegenüber denen diese Eigenschaften invariant sind. Wenn man die Homöomorphien durch eine allgemeinere Klasse von gewissen stetigen Abbildungen \mathfrak{K} ersetzt, so unterscheidet man unter sämtlichen topologischen Eigenschaften eine engere Klasse von den, gegenüber den zu \mathfrak{K} gehörenden Abbildungen invarianten Eigenschaften. Diese Eigenschaften werden wir *\mathfrak{K} -Invarianten* nennen (vgl. [4]). Über die Klasse \mathfrak{K} werden wir nur voraussetzen, dass die Zusammensetzung zweier ihr angehörenden Abbildungen wieder zu ihr gehört.

Wir werden sagen, dass zwei Räume X und Y zum denselben *\mathfrak{K} -Typus* gehören, wenn sie dieselben \mathfrak{K} -Eigenschaften haben. Es ist leicht zu bemerken, dass zwei Räume X und Y dann und nur dann zu demselben \mathfrak{K} -Typus gehören, wenn es zwei \mathfrak{K} -Abbildungen

$$\begin{array}{ccc} f: X & \longrightarrow & Y \text{ und } g: Y \longrightarrow X \\ & \text{auf} & \text{auf} \end{array}$$

gibt.

Betrachten wir einige Beispiele:

1. Es bezeichne \mathfrak{K} die Klasse sämtlicher stetigen Abbildungen. Zu den \mathfrak{K} -Invarianten gehören dann zum Beispiel: Kompaktheit, Separabilität, Zusammenhang und, für Kompakte, auch der lokale Zusammenhang.

2. Viel interessanter ist der Fall, wo \mathfrak{K} die Klasse aller sogenannten *r -Abbildungen* ist. Man versteht dabei unter einer *r -Abbildung* eine stetige Abbildung

$$\begin{array}{ccc} f: X & \longrightarrow & Y \\ & \text{auf} & \end{array}$$

für die eine stetige rechtsseitig inverse Abbildung existiert, dass heisst eine Abbildung g

$$g: Y \underset{\text{in}}{\longrightarrow} X,$$

die der Bedingung $fg(y) = y$ für jeden Punkt $y \in Y$ genügt. Die Klasse der Invarianten von r -Abbildungen ist sehr reich, und somit weisen zwei demselben r -Typus angehörende Räume eine weitreichende Ähnlichkeit ihrer Eigenschaften auf.

Ähnlicherweise kann man die Invarianten von sämtlichen offenen Abbildungen, oder sämtlichen stetigen Abbildungen mit endlichen Urbildmengen, oder sämtlichen stetigen Abbildungen mit azyklischen Urbildmengen und so weiter betrachten. Jeder solchen Invariantenklasse entspricht die Einteilung sämtlicher Räume in entsprechende \mathfrak{R} -Typen.

Da die volle topologische Charakterisierung der Klasse der Polyeder eher hoffnungslos ist, scheint es zweckmässig zu sein, gewisse Charakterisierung von Polyedern vom Standpunkte von verschiedenen \mathfrak{R} -Klassen aus zu betrachten. Vom Standpunkte der r -Invarianten aus lassen sich die Polyeder als metrisierbare, separable Räume durch folgende Bedingungen charakterisieren:

1. *Lokale Kompaktheit.*
2. *Lokale Zusammenziehbarkeit.*
3. *In jedem Punkte eine endliche Dimension.*

Somit können wir diese drei Eigenschaften als die Axiome der Regelmässigkeit, vom Standpunkte der Theorie der r -Invarianten aus, betrachten.

In ähnlicher Weise kann man bei den topologischen Axiomen der dritten Gruppe, anstatt der vollen topologischen Charakterisierung, eine relative Charakterisierung, das heisst eine Charakterisierung im Sinne eines gewissen \mathfrak{R} -Typus verlangen. Man kann erwarten, dass eine systematische Klassifikation der topologischen Invarianten erlauben wird, auf diesem Wege die Lage der klassischen Räume unter den allgemeinen topologischen Räumen klar zu bestimmen.

Bibliographie

- [1] BING, R. H., *Partitioning a set*. Bulletin of the American Mathematical Society, Bd. 55 (1949), S. 1101–1110.
- [2] BIRKHOFF, Garrett, *Metric foundations of geometry I*. Transactions of the American Mathematical Society, Bd. 55 (1944), S. 465–492.
- [3] BLUMENTHAL, L., *Theory and applications of distance geometry*. Oxford 1953, S. 1–347.
- [4] BORSUK, K., *On the topology of retracts*. Annals of Mathematics, Bd. 48 (1947), S. 1082–1094.
- [5] —, *Sur l'élimination de phénomènes paradoxaux en topologie générale*. Proceedings of the International Congress of Mathematicians, Band I, Amsterdam 1954, S. 1–12.
- [6] FRÉCHET, M., *Les espaces abstraits*. Paris 1928, S. XI+296.
- [7] FREUDENTHAL, H., *Neuere Fassungen des Riemann-Helmholtz-Lieschen Raum-problems*. Mathematische Zeitschrift, Bd. 63 (1956), S. 374–405.
- [8] HELMHOLTZ, H., *Über die tatsächliche Grundlagen der Geometrie*. Wissenschaftliche Abhandlungen, Bd. II (1883), S. 610–617.
- [9] HILBERT, D., *Grundlagen der Geometrie*, Leipzig 1900.
- [10] KAMPEN VAN, E. R., *On some characterization of 2-dimensional manifolds*. Duke Mathematical Journal, Bd. 1 (1935), S. 74.
- [11] KOLMOGOROFF, A., *Zur topologisch-gruppentheoretischen Begründung der Geometrie*. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse (1930), S. 208–210.
- [12] KOSIŃSKI, A., *A topological characterization of 2-polytopes*. Bulletin de l'Académie Polonaise des Sciences, Cl. III, Bd. II (1954), S. 321–323.
- [13] KURATOWSKI, C., *Sur l'opération \bar{A} de l'Analysis Situs*. Fundamenta Mathematicae, Bd. 3 (1922), S. 182–199.
- [14] —, *Topologie I*, Monografie Matematyczne, Warszawa 1952, S. XI + 450.
- [15] LIE, S., *Über die Grundlagen der Geometrie*. Gesammelte Abhandlungen II (1922), S. 380–468.
- [16] LINDENBAUM, A., *Contribution à l'étude de l'espace métrique I*. Fundamenta Mathematicae, Bd. 8 (1926), S. 209–222.
- [17] MENGER, K., *Untersuchungen über allgemeine Metrik*. Mathematische Annalen, Bd. 100 (1928), S. 75–163.
- [18] —, *Géométrie générale*. Mémorial des Sciences Mathématiques. Bd. 124, Paris 1954, S. 1–80.
- [19] TANAKA, Tadashi and TOMINAGA, Akira. *Convexification of locally connected generalized continua*. Journal of Science of the Hiroshima University, Bd. 19 (1955), S. 301–306.
- [20] TITS, J., *Etude de certains espaces métriques*. Bulletin de la Société Mathématique de Belgique (1953), S. 44–52.
- [21] —, *Sur un article précédent: Etudes de certaines espaces métriques*. Bulletin de la Société Mathématique de Belgique (1953), S. 124–125.
- [22] TOMINAGA, Akira. *On some properties of non-compact Peano spaces*. Journal of Science of the Hiroshima University, Bd. 19 (1956), S. 457–467.

- [23] URYSOHN, P., *Zum Metrisationsproblem*. Mathematische Annalen, Bd. 94 (1925), S. 309–315.
- [24] VIETORIS, L., *Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen*. Mathematische Annalen, Bd. 97 (1927), S. 454–472.
- [25] WANG, H. C., *Two-point homogeneous spaces*. Annals of Mathematics, Bd. 55(1952), S. 177–191.
- [26] WILSON, W. A., *A relation between metric and euclidean spaces*. American Journal of Mathematics, Bd. 54 (1932), S. 505–517.

LATTICE-THEORETIC APPROACH TO PROJECTIVE AND AFFINE GEOMETRY

BJARNI JÓNSSON

University of Minnesota, Minneapolis, Minnesota, U.S.A.

The results that we are going to discuss are due to several authors. The earliest work along these lines was done by Menger in the late twenties. He was joined a few years later by von Neumann and Birkhoff. A large number of more recent contributions can be found in the papers listed in bibliography; we shall in particular make use of results due to Frink and Schützenberger on projective geometry, and by Croisot, Maeda, Sasaki and Wilcox on affine geometry and its generalizations. The bibliography includes a number of papers that are not concerned directly with geometry, but in which at least some of the ideas and methods were suggested by the investigations of geometric lattices.

1. Concepts from lattice theory. A *lattice* can be defined as a partially ordered set in which any two elements have a least upper bound and a greatest lower bound. We shall use \leq for the partially ordering relation and write $x + y$ and xy for the least upper bound, or sum, and the greatest lower bound, or product, of two elements x and y . Most of our lattices will be *complete*, i.e., any system of elements x_i , $i \in I$, will have a least upper bound and a greatest lower bound

$$\sum_{i \in I} x_i \text{ and } \prod_{i \in I} x_i.$$

In any complete lattice there exist a zero element 0 and a unit element 1 such that $0 \leq x \leq 1$ for every lattice element x . Even when we consider lattices that are not complete we shall always assume that they have a zero element and a unit element. A lattice is said to be *complemented* if for any element x there exists an element y such that $x + y = 1$ and $xy = 0$. If, for any elements a , b , and x with $a \leq x \leq b$ there exists an element y such that $x + y = b$ and $xy = a$, then the lattice is said to be *relatively complemented*. Clearly every relatively complemented lattice (with a zero element and a unit element) is complemented.

An element a is said to *cover* an element b if $b < a$ and if there exists

no element x such that $b < x < a$. An element that covers 0 is called an *atom*, and an element covered by 1 is called a *dual atom*. A lattice in which every element is a sum of atoms is said to be *atomistic*. A system of elements $x_i, i \in I$, in a complete lattice is said to be *independent* if

$$\left(\sum_{i \in J} x_i\right) \left(\sum_{i \in K} x_i\right) = 0$$

whenever J and K are disjoint subsets of I .

We are primarily interested in lattices that are not distributive, but certain special cases of the distributive law will play an important role. A complete lattice is said to be *continuous*¹ if the equation

$$a \sum_{i \in I} x_i = \sum_{i \in I} ax_i$$

holds whenever the set $\{x_i | i \in I\}$ is directed. (A partially ordered set is said to be directed if any two elements of the set have an upper bound that also belongs to the set.) Two elements b and c are said to form a *modular pair* — in symbols $M(b, c)$ — if

$$(x + b)c = x + bc \text{ whenever } x \leq c.$$

If this holds for any two elements b and c , then the lattice is said to be *modular*. If the relation M is symmetric, i.e., if for any two elements b and c the conditions $M(b, c)$ and $M(c, b)$ are equivalent, then the lattice is said to be *semi-modular*. Finally, a lattice is said to be *special* if any two elements that are not disjoint form a modular pair, i.e., if the condition $M(b, c)$ holds whenever $bc \neq 0$.

2. Geometries and geometric lattices. It is convenient for our purpose to take as the undefined concepts of geometry the set consisting of all the points and the function which associates with every set of points the subspace which it spans. Thus we introduce:

DEFINITION 2.1. *By a GEOMETRY we mean an ordered pair $\langle S, C \rangle$ consisting of a set S and a function C which associates with every subset X of S another subset $C(X)$ of S in such a way that the following conditions are*

¹ Such lattices are sometimes called upper continuous, but since the dual concept of a lower continuous lattice will not be needed here, no confusion will be caused by the present terminology.

satisfied:

- (i) $X \subseteq C(X) = C(C(X))$ for every subset X of S .
- (ii) $C(p) = p$ for every $p \in S$ ²
- (iii) $C(\phi) = \phi$ ³
- (iv) For every subset X of S , $C(X)$ is the union of all sets of the form $C(Y)$ with Y a finite subset of X .

DEFINITION 2.2. Suppose $\langle S, C \rangle$ is a geometry.

- (i) An element of S is called a POINT of $\langle S, C \rangle$.
- (ii) A set of the form $C(X)$ with $X \subseteq S$ is called a SUBSPACE of $\langle S, C \rangle$. If $Y = C(X)$, then Y is said to be SPANNED by X .
- (iii) A subspace of $\langle S, C \rangle$ is said to be n -DIMENSIONAL if it is spanned by a set with $n + 1$ elements but is not spanned by any set with fewer than $n + 1$ elements.
- (iv) By a LINE and a PLANE of $\langle S, C \rangle$ we mean, respectively, a one dimensional and a two dimensional subspace of $\langle S, C \rangle$.

From 2.1(iv) it follows that if X and Y are subsets of S , and if $X \subseteq Y$, then $C(X) \subseteq C(Y)$. Together with 2.1(i)–(iii) this yields:

THEOREM 2.3. The family \mathcal{A} of all subspaces of a geometry $\langle S, C \rangle$ has the following properties:

- (i) S and ϕ are members of \mathcal{A} .
- (ii) Every one-element subset of S is a member of \mathcal{A} .
- (iii) The intersection of any number (finite or infinite) of sets belonging to \mathcal{A} is a member of \mathcal{A} .

The tieup between geometries and lattices is now easily established. In fact, if a family \mathcal{A} of subsets of a set S has the properties 2.3(i)–(iii), then \mathcal{A} is a complete and atomistic lattice under set-inclusion. The lattice product of any system X_i , $i \in I$, of sets belonging to \mathcal{A} is their set-theoretic intersection, and the lattice sum of the sets X_i is the smallest member of \mathcal{A} which contains their union. The atoms of \mathcal{A} are the one-element subsets of S . Conversely, any complete and atomistic lattice A is isomorphic to a family \mathcal{A} consisting of subsets of some set S and satisfying 2.3(i)–(iii). In fact, we may take for S the set of all atoms of A and

² Strictly speaking $C(\{p\}) = \{p\}$. We shall also write $C(p, q)$, $C(p, q, r)$, ..., $C(X, p)$, $C(X, p, q)$, ... for $C(\{p, q\})$, $C(\{p, q, r\})$, ..., $C(X \cup \{p\})$, $C(X \cup \{p, q\})$, ...

³ ϕ is the empty set.

correlate with each element x of A the set consisting of all the atoms p of A for which $p \leq x$. This leads to

DEFINITION 2.4. *A lattice is said to be GEOMETRIC if it is isomorphic to the lattice of all subspaces of some geometry.*

Each of the next two theorems gives an axiomatic characterization of geometric lattices. The first is an immediate consequence of the definitions involved.

THEOREM 2.5. *A lattice is geometric if and only if it is complete and atomistic, and has the property that for any atom p and any systems of atoms q_i , $i \in I$, the condition*

$$p \leq \sum_{i \in I} q_i$$

implies that there exists a finite subset J of I such that

$$p \leq \sum_{i \in J} q_i$$

THEOREM 2.6. *A lattice is geometric if and only if it is complete, atomistic and continuous.*

3. The exchange property. Our notion of a geometry is an extremely general one and cannot be expected to have many interesting consequences. It may for instance happen that two distinct lines have more than one point in common, and in fact it is easy to construct geometries where one line is properly contained in another. We now consider a condition which excludes such pathological situations.

DEFINITION 3.1. *A geometry $\langle S, C \rangle$ is said to have the EXCHANGE PROPERTY if, for any points p and q and any subset X of S , the conditions $p \in C(X, q)$ and $p \notin C(X)$ jointly imply that $q \in C(X, p)$.*

DEFINITION 3.2. *By a MATROID LATTICE we mean a geometric lattice with the property that, for any atoms p and q and any element x , the conditions $p \leq q + x$ and $p \not\leq x$ jointly imply that $q \leq p + x$.*

THEOREM 3.3. *In order for a lattice A to be isomorphic to the lattice of all subspaces of a geometry which has the exchange property it is necessary and sufficient that A be a matroid lattice.*

THEOREM 3.4. *Every matroid lattice is relatively complemented.*

Matroid lattices have been extensively investigated. Of the numerous equivalent characterizations of this class of lattices, the one given in the next theorem is particularly interesting.

THEOREM 3.5. *In order for a lattice A to be a matroid lattice it is necessary and sufficient that A be complete, atomistic, continuous and semi-modular.*

THEOREM 3.6. *In any matroid lattice the following conditions hold for all elements a, b, c, d , and all atoms $p, q, p_0, p_1, \dots, p_n$:*

- (i) *If $a < a + p \leq a + q$, then $a + p = a + q$.*
- (ii) *If $ap = 0$, then $a + p$ covers a .*
- (iii) *If $(a + b)p = 0$, then $(a + p)b = ab$.*
- (iv) *If $(p_0 + p_1 + \dots + p_{k-1})p_k = 0$ for $k = 1, 2, \dots, n$, then the system $p_i, i = 0, 1, \dots, n$, is independent.*
- (v) *If a and b cover ab , then $a + b$ covers a and b .*
- (vi) *If a covers ab , then $a + b$ covers b .*
- (vii) *If b covers bc , then $M(b, c)$.*
- (viii) *If $bc < a < c < b + c$, then there exists an element x such that $bc < x \leq b$ and $a = (a + x)c$.*
- (ix) *If $bc < a < c < b + c$, then there exists an element x such that $bc < x \leq b$ and $(a + x)c < c$.*
- (x) *If $bc < a < c < a + b$, then there exists an element x such that $bc < x \leq b$ and $a = (a + x)c$.*

Conversely, any geometric lattice which satisfies one of the conditions (i)–(x) is a matroid lattice.

4. Strongly planar geometries. In the classical approach to affine geometry, a set X of points is by definition a subspace if and only if it contains every line with which it has two distinct points in common, and contains every plane with which it has three non-collinear points in common. In projective geometry the first of these two conditions alone is taken as the characteristic property of a subspace. Thus it is true in either case that if $C(p, q, r) \subseteq X$ whenever $p, q, r \in X$, then X is a subspace. Another property common to the classical affine and projective geometries is the fact that two intersecting planes which are contained in the same 3-space have a line in common. This motivates the next two definitions.

DEFINITION 4.1. A geometry $\langle S, C \rangle$ is said to be **PLANAR** if it has the exchange property and, for every subset X of S , the condition

$$C(p, q, r) \subseteq X \text{ whenever } p, q, r \in X$$

implies that X is a subspace of $\langle S, C \rangle$.

DEFINITION 4.2. A geometry $\langle S, C \rangle$ is said to be **STRONGLY PLANAR** if it is planar and has the property that any two distinct planes that are contained in the same 3-space are either disjoint or else their intersection is a line.

No simple condition is known which characterizes those lattices which correspond to planar geometries. As regards strongly planar geometries we have:

THEOREM 4.3. A geometry $\langle S, C \rangle$ is strongly planar if and only if it has the exchange property and, for any points p, q, r , and any set X of points, the conditions

$$p \in C(X, q), \quad r \in C(X)$$

jointly imply that there exists a point s such that

$$p \in C(q, r, s) \text{ and } s \in C(X).$$

THEOREM 4.4. For any matroid lattices A the following conditions are equivalent:

- (i) A is isomorphic to the lattice of all subspaces of a strongly planar geometry.
- (ii) For any atoms p, q, r of A and any element a of A , the conditions

$$p \leq q + a \text{ and } r \leq a$$

jointly imply that there exists an atom s such that

$$p < q + r + s \text{ and } s \leq a.$$

- (iii) A is special.
- (iv) For any element a of A and any dual atom h of A , if $0 < ah < a$, then a covers ah .

5. Projective geometries. In order to obtain a concept that corresponds more or less to the classical notion of a projective geometry we need axioms to the effect that any two lines in the same plane have a point in common, and that if a set X of points has the property that it contains every line with which it has two points in common, then X is a subspace. These two conditions can be stated as a single axiom:

DEFINITION 5.1. A geometry $\langle S, C \rangle$ is said to be PROJECTIVE if it has the exchange property and, for any points p and q and any set X of points, the conditions

$$p \in C(X, q) \text{ and } p \neq q$$

jointly imply that there exist a point r such that

$$p \in C(q, r) \text{ and } r \in C(X).$$

DEFINITION 5.2. A lattice is said to be PROJECTIVE if and only if it is isomorphic to the lattice of all subspaces of a projective geometry.

COROLLARY 5.3. Every projective geometry is strongly planar.

COROLLARY 5.4. Suppose p is a point and X and Y are sets of points of a projective geometry $\langle S, C \rangle$. If

$$p \in C(X, Y), \quad p \notin C(X) \text{ and } p \notin C(Y),$$

then there exist points q and r such that

$$p \in C(q, r), \quad q \in C(X) \text{ and } r \in C(Y).$$

With the aid of these two corollaries we get a particularly elegant characterization of projective lattices:

THEOREM 5.5. A lattice is projective if and only if it is complete, atomistic, continuous and modular.

The notion of a projective geometry as defined here is obviously more general than the classical concept, since we put no restriction on the dimension and do not exclude geometries in which there are degenerate lines consisting of only two distinct points. However, this generalization is less radical than it might appear at first glance. Every projective lattice A is a direct product of indecomposable sublattices,

$$A = \prod_{i \in I} A_i.$$

When applied to the lattice of all subspaces of a projective geometry $\langle S, C \rangle$, this decomposition corresponds to a partitioning of S into subspaces S_i in such a way that two distinct points belong to the same subspace if and only if they determine a non-degenerate line. Some of these components may be trivial, consisting of just one point or of just one line, and others may be non-Arguesian planes. With these exceptions, we can associate with each component S_i a division ring and introduce coordinates

in the manner of classical geometry, with the sole difference that the number of coordinates may be infinite.

This brings us to the subject of Desargues' Law:

DEFINITION 5.6. A geometry $\langle S, C \rangle$ is said to be ARGUESIAN if it is projective and, for any points $p_0, p_1, p_2, q_0, q_1, q_2$, the condition

$$C(p_1, q_1) \cap C(p_2, q_2) \subseteq C(p_0, q_0)$$

implies that

$$C(p_1, p_2) \cap C(q_1, q_2) \subseteq C((C(p_0 p_1) \cap C(q_0, q_1)) \cup (C(p_0, p_2) \cap C(q_0, q_2))).$$

DEFINITION 5.7. A lattice is said to be ARGUESIAN if and only if it is isomorphic to the lattice of all subspaces of an Arguesian geometry.

The formulation of Desargues' Law in Definition 5.6 differs from the classical version in that no restriction is placed on the six points involved (such as that they be distinct, or that the three pairs $p_i, q_i, i = 0, 1, 2$ lie on three distinct but concurrent lines). However, the two formulations are actually equivalent, for some of the special cases that are normally excluded are actually valid in all projective geometries, while the remaining cases follow from the classical Desargues' Law.

It is of course easy to write down a lattice-theoretic version of Desargues' Law, involving six atoms (5.8(ii)). It is an interesting fact that this condition actually holds with the six atoms replaced by any six lattice elements (5.8(iii)). Perhaps more important, however, is the fact that this condition is actually equivalent to a lattice identity (5.8(iv)).

THEOREM 5.8. If A is a geometric lattice, then the following conditions are equivalent:

- (i) A is Arguesian
- (ii) A is modular and, for any atoms $p_0, p_1, p_2, q_0, q_1, q_2$ of A , the condition

$$(p_1 + q_1)(p_2 + q_2) \leq p_0 + q_0$$

implies that

$$(p_1 + p_2)(q_1 + q_2) \leq (p_0 + p_1)(q_0 + q_1) + (p_0 + p_2)(q_0 + q_2)$$

- (iii) For any elements $a_0, a_1, a_2, b_0, b_1, b_2 \in A$, the condition

$$(a_1 + b_1)(a_2 + b_2) \leq a_0 + b_0$$

implies that

$$(a_1 + a_2)(b_1 + b_2) \leq (a_0 + a_1)(b_0 + b_1) + (a_0 + a_2)(b_0 + b_2).$$

(iv) For any elements $a_0, a_1, a_2, b_0, b_1, b_2 \in A$, if

$$y = (a_1 + a_2)(b_1 + b_2)[(a_0 + a_1)(b_0 + b_1) + (a_0 + a_2)(b_0 + b_2)].$$

then

$$(a_0 + b_0)(a_1 + b_1)(a_2 + b_2) \leq a_0(a_1 + y) + b_0(b_1 + y).$$

Observe that in (iii) and (iv) we do not assume the modular law; it turns out to be a consequence of the given conditions.⁴ In terms of the decomposition discussed above, a projective lattice A is Arguesian if and only if none of its indecomposable factors is isomorphic to the lattice of all subspaces of a non-Arguesian projective plane.

6. Affine geometries. We define an affine geometry to be a strongly planar geometry in which Euclid's parallel axiom holds:

DEFINITION 6.1. A geometry $\langle S, C \rangle$ is said to be AFFINE if and only if $\langle S, C \rangle$ is strongly planar and has the following property: For any plane P , line L , and point p , the conditions

$$p \in P, L \subseteq P \text{ and } p \notin L$$

jointly imply that there exists a unique line L' such that

$$p \in L', L' \subseteq P \text{ and } L \cap L' = \phi.$$

DEFINITION 6.2. A lattice is said to be AFFINE if and only if it is isomorphic to the lattice of all subspaces of an affine geometry.

THEOREM 6.3. A lattice A is affine if and only if it is a special matroid lattice with the following property: For any atoms p, q , and r , if

$$p < p + q < p + q + r,$$

then there exists a unique element x such that

$$r < x < p + q + r \text{ and } (p + q)x = 0$$

The relation between our concepts of an affine geometry and of a non-degenerate projective geometry is precisely analogous to the relation between their classical counterparts:

⁴ In fact, in any lattice A , (iv) implies (iii) and (iii) in turn implies that A is modular.

THEOREM 6.4. *If p is an atom of an affine lattice A , then the set of all elements $x \in A$ with $p \leq x$ is an indecomposable projective lattice under the partially ordering relation defined on A .*

THEOREM 6.5. *If h is a dual atom of an indecomposable projective lattice A , then the set A_h consisting of 0 and of all elements $x \in A$ with $x \leq h$ is an affine lattice under the partially ordering relation on A . Conversely, for any affine lattice B there exist an indecomposable projective lattice A and a dual atom h of A such that B is isomorphic to A_h .*

7. Applications of geometry to lattice theory. As can be seen from the above discussion, the applications of lattice theory to the axiomatization of geometry have yielded radically different and quite simple characterizations of the geometries considered. Similar work has been done with other types of geometries, and it is quite certain that more can be done along these lines.

But these investigations have also aided, both directly and indirectly, in the study of certain problems in lattice theory. We shall mention briefly some examples that illustrate this point.

Modular lattices may be regarded as a generalization of projective geometry. Since every projective lattice is complemented, it might be more reasonable to consider only complemented modular lattices. Just how far-reaching a generalization is this? A partial answer was provided by von Neumann, who showed that every complemented modular lattice which satisfies certain conditions (namely, possesses an n -frame with $n \geq 4$) is isomorphic to the lattice of all principal left ideals of a regular ring. Since a full matrix ring over a division ring is regular, this may be regarded as a generalization of the coordinatization theorem for non-Arguesian geometries. There is however an important problem open here: To find a condition that is both necessary and sufficient in order for a complemented modular lattice to be isomorphic to the lattice of all principal left ideals of a regular ring.

A representation of a different kind was obtained by Frink, who proved that every complemented modular lattice B is a sublattice of a projective lattice A . The Frink geometry associated with B is a generalization of the Stone space of a Boolean algebra; its points are the maximal proper dual ideals of B , and the line through two points P and Q consists of all points R such that $P \cap Q \subseteq R$. The subspace correlated with a given element $x \in B$ is the set of all points P such that $x \in P$. It is known that this

embedding preserves all identities that hold in B , and from this it follows in particular that A is Arguesian if and only if B satisfies the condition (iv) of Theorem 5.8.

Still other investigations, by Bear and Inaba, that were inspired by the coordinatization theorem, concern the lattice of all submodules of a module over a ring of a certain type and, as a special case, the lattices of all subgroups of a finite Abelian group. The principal result may be regarded as a representation theorem for a certain class of modular lattices, including all the finite dimensional projective lattices.

Even for modular lattices that are not complemented, Desargues' Law in the form of the condition (iv) of Theorem 5.8 turns out to be significant. Most of the modular lattices that arise in applications of lattice theory are isomorphic to lattices of commuting equivalence relations, and in fact all the known examples for which this is not the case are of a somewhat pathological character. It is therefore natural to try to characterize axiomatically the class of all those lattices for which such a representation exists. It is not hard to prove that Desargues' Law is a necessary condition, but it is still an open question whether this is also sufficient. On the other hand, an infinite system of axioms (in the form of conditional equations) is known, which is sufficient as well as necessary, and these axioms are such that when applied to the lattice of all subspaces of a projective geometry, they reduce to certain configuration theorems which are valid in all Arguesian geometries.

The family of all equivalence relations over a set U , or equivalently the family of all partitions of U , is a geometric lattice. The class consisting of all lattices of this form (and of their isomorphic images) can be conveniently characterized by describing the corresponding geometries. In fact, in order for the lattice of all subspaces of a geometry $\langle S, C \rangle$ to be isomorphic to the lattice of all equivalence relations over some set, it is necessary and sufficient that the following conditions be satisfied:

- (1) $\langle S, C \rangle$ is planar and has the exchange property.
- (2) Each plane of $\langle S, C \rangle$ has either 3, 4, or 6 points.
- (3) For each line L of $\langle S, C \rangle$, either L has exactly two points, and there are exactly two lines parallel to L , or else L has exactly three points and there is no line parallel to L .

In Theorem 6.5, affine lattices are characterized as those lattices which can be obtained from indecomposable projective lattices by removing all the elements x contained in some fixed dual atom h , with the exception of the zero element. If h is not a dual atom, this process still leads to a special

matroid lattice, but only half of Euclid's parallel axiom will be satisfied (the uniqueness part). The question of what lattices can be obtained from complemented modular lattices by removing more general sets, subject to some suitable conditions, has been studied by Wilcox. Some of his results have been announced in abstracts, but a detailed account has not yet appeared.

These examples will suffice to illustrate the fact that the investigations of the connections between geometries and lattices have yielded something of interest to both subjects.

Bibliography

- AMEMIYA, I. *On the representation of complemented modular lattices*. Journal of the Mathematical Society of Japan, vol. 9 (1957), pp. 263–279.
- BAER, R., *A unified theory of projective spaces and finite abelian groups*. Transactions of the American Mathematical Society, vol. 52 (1942), pp. 283–343.
- , *Linear algebra and projective geometry*. New York, 1952, VIII + 318 pp.
- BIRKHOFF, G., *Abstract linear dependence and lattices*. American Journal of Mathematics, vol. 57 (1935), pp. 800–804.
- , *Combinatory relations in projective geometry*. Annals of Mathematics (2), vol. 36 (1935), pp. 743–748.
- , *Lattice theory*. New York 1948, XIII + 283 pp.
- , *Metric foundations of geometry I*. Transactions of the American Mathematical Society, vol. 55 (1944), pp. 465–492.
- , and FRINK, O., *Representations of lattices by sets*. Transactions of the American Mathematical Society, vol. 64 (1948), pp. 299–316.
- and VON NEUMANN, J., *The logic of quantum mechanics*. Annals of Mathematics (2), vol. 37 (1936), pp. 823–843.
- CROISOT, R., *Axiomatique des treillis semi-modulaires*. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), vol. 231 (1950), pp. 12–14.
- , *Contribution à l'étude des treillis semi-modulaires de longueur infinie*. Annales Scientifiques de l'École Normale Supérieure (3), vol. 68 (1951), pp. 203–265.
- , *Diverses caractérisations des treillis semi-modulaires, modulaires et distributifs*. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), vol. 231 (1950), pp. 1399–1401.
- , *Quelques applications et propriétés des treillis semi-modulaires de longueur infinie*. Annales de la Faculté des Sciences de l'Université de Toulouse pour les Sciences Mathématiques et les Sciences Physiques (4), vol. 16 (1952) pp. 11–74.
- , *Sous-treillis, produit cardinaux et treillis homomorphes des treillis semi-modu-*

- laire*. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), vol. 232 (1951), pp. 27–29.
- DILWORTH, R. P., *Dependence relations in a semi-modular lattice*. Duke Mathematical Journal, vol. 11 (1944), pp. 575–587.
- , *Ideals in Birkhoff lattices*. Transactions of the American Mathematical Society, vol. 49 (1941), pp. 325–353.
- , *Note on complemented modular lattices*. Bulletin of the American Mathematical Society, vol. 46 (1940), pp. 74–76.
- , *The arithmetic theory of Birkhoff lattices*. Duke Mathematical Journal, vol. 8 (1941), pp. 286–299.
- DUBREIL-JACOTIN, M. L., LESIEUR, L. and CROISOT, R., *Leçons sur la théorie des treillis des structures algébriques ordonnées et des treillis géométriques*. Paris 1953, VIII + 385 pp.
- FICKEN, F. A., *Cones and vector spaces*. American Mathematical Monthly, vol. 47 (1940), pp. 530–533.
- FRINK, O., Jr., *Complemented modular lattices and projective spaces of infinite dimension*. Transactions of the American Mathematical Society, vol. 60 (1946), pp. 452–467.
- FRYER, K. D. and HALPERIN, I., *Coordinates in geometry*. Transactions of the Royal Society of Canada, vol. 48 (1954), pp. 11–26.
- , and —, *On the coordinatization theorem of J. von Neumann*. Canadian Journal of Mathematics, vol. 7 (1955), pp. 432–444.
- and —, *The von Neumann coordinatization theorem for complemented modular lattices*. Acta Universitatis Szegediensis. Acta Scientiarum Mathematicarum, vol. 16 (1956), pp. 203–249.
- HALL, M. and DILWORTH, R. P., *The imbedding problem for modular lattices*. Annals of Mathematics (2), vol. 45 (1944), pp. 450–456.
- HALPERIN, I., *Additivity and continuity of perspectivity*. Duke Mathematical Journal, vol. 5 (1939), pp. 503–511.
- , *Dimensionality in reducible geometries*. Annals of Mathematics (2), vol. 40 (1939), pp. 581–599.
- , *On the transitivity of perspectivity in continuous geometries*. Transactions of the American Mathematical Society, vol. 44 (1938), pp. 537–562.
- Hsu, C., *On lattice theoretic characterization of the parallelism in affine geometry*. Annals of Mathematics (2) vol. 50 (1949), pp. 1–7.
- INABA, E., *On primary lattices*. Journal of the Faculty of Science, Hokkaido University, vol. 11 (1948), pp. 39–107.
- , *Some remarks on primary lattices*. Natural Science Report of the Ochanomizu University, vol. 2 (1951), pp. 1–5.
- IWAMURA, T., *On continuous geometries*. I. Japanese Journal of Mathematics, vol. 19 (1944), pp. 57–71.
- , *On continuous geometries*. II. Journal of the Mathematical Society of Japan. vol. 2 (1950), pp. 148–164.
- IZUMI, S., *Lattice theoretic foundation of circle geometry*. Proceedings of the Imperial Academy (Tokyo), vol. 16 (1940), pp. 515–517.
- JÓNSSON, B., *Modular lattices and Desargues' theorem*. Mathematica Scandinavica, vol. 2 (1954), pp. 295–314.

- , *On the representation of lattices*. *Mathematica Scandinavica*, vol. 1 (1953), pp. 193–206.
- KAPLANSKY, I., *Any orthocomplemented complete modular lattice is a continuous geometry*. *Annals of Mathematics* (2), vol. 61 (1955), pp. 524–541.
- KODAIRA, K. and HURUYA, S., *On continuous geometries I, II, III* (in Japanese). *Zenkoku Shijo Sukaku Danwakai*, vol. 168 (1938), pp. 514–531; vol. 169 (1938), pp. 593–609; vol. 170 (1938), pp. 638–656.
- KÖTHE, G., *Die Theorie der Verbände, ein neuer Versuch zur Grundlegung der Algebra und der projectiven Geometrie*. *Jahresbericht der Deutschen Mathematiker Vereinigung*, vol. 47 (1937), pp. 125–144.
- KRISHNAN, V. S., *Partially ordered sets and projective geometry*. *The Mathematics Student*, vol. 12 (1944), pp. 7–14.
- LOOMIS, I. H., *The lattice theoretic background of the dimension theory of operator algebras*. *Memoirs of the American Mathematical Society* 1955, No. 18, 36 pp.
- MACLANE, S., *A lattice formulation for transcendence degrees and p -bases*. *Duke Mathematical Journal*, vol. 4 (1938), pp. 455–468.
- MAEDA, F., *A lattice formulation for algebraic and transcendental extensions in abstract algebras*. *Journal of Science of the Hiroshima University*, vol. 16 (1952–1953), pp. 383–397.
- , *Continuous geometry* (in Japanese). Tokyo 1952, 2 + 3 + 225 pp.
- , *Dimension functions on certain general lattices*. *Journal of Science of the Hiroshima University*, vol. 19 (1955), pp. 211–237.
- , *Dimension lattice of reducible geometries*. *Journal of Science of the Hiroshima University*, vol. 13 (1944), pp. 11–40.
- , *Direct sums and normal ideals of lattices*. *Journal of Science of the Hiroshima University*, vol. 14 (1949–1950), pp. 85–92.
- , *Embedding theorem of continuous regular rings*. *Journal of Science of the Hiroshima University*, vol. 14 (1949–1950), pp. 1–7.
- , *Lattice theoretic characterization of abstract geometries*. *Journal of Science of the Hiroshima University*, vol. 15 (1951–1952), pp. 87–96.
- , *Matroid lattices of infinite length*. *Journal of Science of the Hiroshima University*, vol. 15 (1951–1952), pp. 177–182.
- , *Representations of orthocomplemented modular lattices*. *Journal of Science of the Hiroshima University*, vol. 14 (1949–1950), pp. 93–96.
- , *The center of lattices*. (In Japanese). *Journal of Science of the Hiroshima University*, vol. 12 (1942), pp. 11–15.
- MENGER, K., *Algebra der Geometrie (Zur Axiomatik der projectiven Verknüpfungsbeziehungen)*. *Ergebnisse eines mathematischen Kolloquiums*, vol. 7 (1936), pp. 11–12.
- , *Axiomatique simplifiée de l'algèbre de la géométrie projective*. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, vol. 206 (1938), pp. 308–310.
- , *Bemerkungen zu Grundlagenfragen IV. Axiomatik der endlichen Mengen und der elementargeometrischen Verknüpfungsbeziehungen*. *Jahresbericht der Deutschen Mathematiker Vereinigung*, vol. 37 (1928), pp. 309–325.
- , *La géométrie axiomatique de l'espace projectif*, *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, vol. 228 (1949), pp. 1273–1274.

- , *New foundations of projective and affine geometry. Algebra of geometry.* Annals of Mathematics (2), vol. 37 (1936), pp. 456–482.
- , *Non-Euclidean geometry of joining and intersecting.* Bulletin of the American Mathematical Society, vol. 44 (1938), pp. 821–824.
- , *On algebra of geometry and recent progress in non-Euclidean geometry.* The Rice Institute Pamphlets, vol. 27 (1940), pp. 41–79.
- , *Selfdual postulates in projective geometry.* American Mathematical Monthly, vol. 55 (1948), p. 195.
- MOUSINHO, M. L., *Modular and projective lattices.* Summa Brasiliensis Mathematicae, vol. 2 (1950), pp. 95–112.
- VON NEUMANN, J., *Algebraic theory of continuous geometries.* Proceedings of the National Academy of Science, U.S.A., vol. 23 (1937), pp. 16–22.
- , *Continuous geometry.* Proceedings of the National Academy of Science, U.S.A., vol. 22 (1936), pp. 92–100.
- , *Continuous rings and their arithmetics.* Proceedings of the National Academy of Science, U.S.A., vol. 23 (1937), pp. 341–349. Errata, *ibid.*, p. 593.
- , *Examples of continuous geometries.* Proceedings of the National Academy of Science, U.S.A. vol. 22 (1936), pp. 101–108.
- , *Lectures on continuous geometries, I–III, Princeton 1936–1937.* (Mimeographed lecture notes.)
- , *On regular rings.* Proceedings of the National Academy of Science, U.S.A. vol. 22 (1936), pp. 707–712.
- , and HALPERIN, I., *On the transivity of perspective mappings.* Annals of Mathematics (2), vol. 41 (1940), pp. 87–93.
- PRENOWITZ, WALTER, *Total lattices of convex sets and of linear spaces.* Annals of Mathematics (2), vol. 49 (1948), pp. 659–688.
- SASAKI, U., *Lattice theoretical characterization of an affine geometry of arbitrary dimension.* Journal of Science of the Hiroshima University, vol. 16 (1952–1953), pp. 223–238.
- , *Lattice theoretic characterization of geometries satisfying "Axiome der Verknüpfung".* Journal of Science of the Hiroshima University, vol. 16 (1952–1953), pp. 417–423.
- , *Orthocomplemented lattices satisfying the exchange axiom.* Journal of Science of the Hiroshima University, vol. 17 (1953–1954), pp. 293–302.
- , *Semi-modularity in relatively atomic, upper continuous lattices.* Journal of Science of the Hiroshima University, vol. 16 (1952–1953), pp. 409–416.
- , and FUJIWARA, S., *The characterization of partition lattices.* Journal of Science of the Hiroshima University, vol. 15 (1951–1952), pp. 189–201.
- and —, *The decomposition of matroid lattices.* Journal of Sciences of the Hiroshima University, vol. 15 (1951–1952), pp. 183–188.
- SCHÜTZENBERGER, M., *Sur certains axiomes de la théorie des structures.* Comptes Rendus Hebdomadaire des Séances de l'Académie des Sciences (Paris), vol. 221 (1945), pp. 218–220.
- WHITNEY, H., *On the abstract properties of linear dependence.* American Journal of Mathematics, vol. 57 (1935), pp. 509–533.
- WILCOX, L. R., *An imbedding theorem for semi-modular lattices.* Bulletin of the American Mathematical Society, vol. 60 (1954), p. 532.

- , *Modular extensions of semi-modular lattices*. Bulletin of the American Mathematical Society, vol. 61 (1955), pp. 524–525.
- , *Modularity in Birkhoff lattices*. Bulletin of the American Mathematical Society, vol. 50 (1944), pp. 135–138.
- , *Modularity in the theory of lattices*. Annals of Mathematics (2), vol. 40 (1939), pp. 490–505.

CONVENTIONALISM IN GEOMETRY *

ADOLF GRÜNBAUM

Lehigh University, Bethlehem, Pennsylvania, U.S.A.

1. **Introduction.** In what sense and to what extent can the ascription of a particular metric geometry to physical space be held to have an *empirical* warrant? To answer this question we must inquire whether and how empirical facts function restrictively so as to support a unique metric geometry as the true description of physical space.

The inquiry is prompted by the conflict of ideas on this issue emerging in the Albert Einstein volume in Schilpp's Library of Living Philosophers between Robertson, Reichenbach and Einstein. Robertson characterizes K. Schwarzschild's attempt to determine *observationally* the Gaussian curvature of an astronomical 2-flat as an inspiring implementation of the *empiricist* conception of physical geometry. And Robertson deems Schwarzschild's view to be "in refreshing contrast to the pontifical pronouncement of Henri Poincaré," [25, p. 325] who had declared that "Euclidean geometry has, . . . , nothing to fear from fresh experiments" [20, p. 81] after reviewing the various possible results of stellar parallax measurements. In the same volume [21, p. 297] and elsewhere [22, Ch. 8; 23, pp. 30-37], Reichenbach maintains, as Carnap had done in his early monograph *Der Raum* [3], that the question as to which metric geometry prevails in physical space is indeed *empirical* but subject to an important proviso: it becomes empirical only *after* a physical definition of congruence for line segments has been given *conventionally* by *stipulating* (to within a constant factor depending on the choice of unit) what length is to be assigned to a transported solid rod in different positions of space. Reichenbach calls this *qualified* empiricist conception "the relativity of geometry" and terms "conventionalism" the more radical thesis that even *after* the physical meaning of "congruent" has been fixed, it is *entirely* a matter of convention which physical geometry is said to prevail. Believing Poincaré to have been an exponent of conventionalism in *this* sense, Reichenbach rejects Poincaré's supposed philosophy of geometry as

* The author is indebted to the National Science Foundation of the U.S.A. for the support of research.

erroneous. On the other hand, Einstein criticizes Reichenbach's relativity of geometry by upholding a particular version of conventionalism which he attributes to Poincaré [9, pp. 676–679].

This exchange reveals that there are several *different* theses concerning the presence of stipulational ingredients in physical geometry and the warrant for their introduction which require critical examination in the course of our inquiry.

Our main concern is with the respective roles of convention and fact in the ascription of a particular metric geometry to physical space on the basis of measurements with a *rigid body*. Accordingly, we shall discuss in turn the two principal problems which have been posed in connection with the formulation of the criterion of rigidity and of isochronism.

2. The Criterion of Rigidity: I. The Status of Spatial Congruence.

Differential geometry allows us to metrize a given physical surface, say an infinite blackboard or some portion of it, in various ways so as to acquire any metric geometry compatible with its topology. Thus, if we have such a space and a net-work of Cartesian coordinates on it, we can just as legitimately metrize the portion *above* the x -axis by means of the metric $ds^2 = \frac{dx^2 + dy^2}{y^2}$, which confers a hyperbolic geometry on

that space, as by the Euclidean metric $ds^2 = dx^2 + dy^2$. The geometer is not disconcerted by the fact that in the former metrization, the lengths of horizontal segments whose termini have the same coordinate differences

dx will be $ds = \frac{dx}{y}$ and will thus depend on where they are along the

y -axis. What is his sanction for preserving equanimity in the face of the fact that this metrization commits him to regard a segment for which $dx = 2$ at $y = 2$ as *congruent* to a segment for which $dx = 1$ at $y = 1$, although the *customary* metrization would regard the length ratio of these segments to be 2 : 1? His answer would be that unless one of two segments is a subset of the other the congruence of two segments is a matter of convention, stipulation or definition and *not* a factual matter concerning which empirical findings could show one to have been mistaken. He does *not* say, of course, that a transported solid rod will coincide successively with the two hyperbolically-congruent segments but allows for this non-coincidence by making the length of the transported rod a suitable function of its position rather than a constant. And in this way, he justifies his claim that the hyperbolic metrization possesses

both epistemological and mathematical credentials as good as those of the Euclidean one.

This conception of congruence was vigorously contested by Bertrand Russell and defended by Poincaré in a controversy which grew out of the publication of Russell's *Foundations of Geometry* [28]. Our first concern will be with the central issue of that debate.

Russell states the factualist's argument as follows [26, pp. 687–688] ¹:

"It seems to be believed that since measurement is necessary to *discover* equality or inequality, these cannot exist without measurement. Now the proper conclusion is exactly the opposite. Whatever one can discover by means of an operation must exist independently of that operation: America existed before Christopher Columbus, and two quantities of the same kind must *be* equal *or* unequal before being measured. Any method of measurement is good or bad according as it yields a result which is true or false. Mr. Poincaré, on the other hand, holds that measurement creates equality and inequality. It follows [then] . . . that there is nothing left to measure and that equality and inequality are terms devoid of meaning."

Before setting forth the grounds for regarding Russell's argument here as untenable, it will be useful to analyze the reasoning employed in an *inadequate* criticism of it. This analysis will exhibit an important facet of the relation of the axiomatic method in pure geometry to the description of physical space.

We are told that Russell's contention can be dismissed by simply pointing to the theory of models: since physical geometry is a semantically-interpreted abstract calculus, the *customary* physical interpretation of the abstract relation term "congruent" (for line segments) as opposed to the kind of interpretation given in our hyperbolic metrization above clearly cannot itself be a factual statement. Hence it is argued that the alternative metrization of spatial and temporal continua should never have been either startling or a matter for dispute. On this view, Poincaré could have spared himself the trouble of polemicizing against Russell on behalf of it in the form of a philosophical doctrine of congruence. For, so the argument runs [7, pp. 9–10], there can be nothing particularly problematic about the physical interpretation of the term "congruent": like the physical meaning of all other primitives of the calculus, the denotata of the abstract relation term "*congruent*" (for line segments) are specified by

¹ An implicit endorsement of this argument is given by H. von Helmholtz [33, p. 15].

semantical rules which are fully on a par in regard to *both* conventionality *and* importance with those furnishing the interpretation of any of the other abstract primitives of the calculus. In fact, Tarski's axioms for elementary Euclidean geometry, which appear in this volume, even dispense with the primitive "congruent" for line segments and yet yield (the elementary form of) a metric geometry by using instead a quaternary predicate ∂ denoting the equidistance relation between 4 points.

That such an argument does not go to the heart of the issue and hence would have failed to convince Russell can be seen from the following: The congruence relation for line segments, and correspondingly for regions of surfaces and of 3-space, is a reflexive, symmetrical and transitive relation in these respective classes of geometrical configurations. Thus, congruence is a kind of *equality* relation. Now suppose that one believes, as Russell and Helmholtz thought they could believe justifiably, that the spatial equality obtaining between congruent line segments consists in their each containing the *same intrinsic* amount of space. Then one will maintain that in any *physico-spatial* interpretation of an abstract geometrical calculus, it is *never* legitimate to choose arbitrarily what specific line segments are going to be called "congruent". And, by the same token, one will assert that in Tarski's aforementioned axiomatization, it is never arbitrary what quartets of physical points are to be regarded as the denotata of his quaternary equidistance predicate ∂ . Instead the imputation of an *intrinsic metric* to the extended continua of space and time will issue in the following contentions: (i) since only "truly equal" intervals may be called "congruent", Newton [18, pp. 6-8] was right in insisting that there is only one true metrization of the time continuum, and (ii) there is no room for choice as to the lines which are to be called "straight" and hence no choice among alternative metric geometries of physical space, since the geodesic requirement $\delta f ds = 0$, which must be satisfied by the straight lines, is imposed subject to the restriction that only *intrinsically congruent* line elements may be assigned the same length ds .

These considerations show that it will *not* suffice in this context simply to take the model-theoretic conception of geometry for granted and thereby to dismiss the Russell-Helmholtz claim peremptorily in favor of alternative metrization. Rather what is needed is a *refutation* of the Russell-Helmholtz *root-assumption* of an *intrinsic metric*: to exhibit the untenability of that assumption is to provide the *justification* of the model-theoretic affirmation that a given set of physico-spatial facts may be held

to be as much a realization of a Euclidean calculus as of a *non*-Euclidean one yielding the same topology.

We shall now see how Riemann and Poincaré furnished the philosophical underpinning for that affirmation.

The following statement in Riemann's Inaugural Dissertation [24, pp. 274, 286] contains a fundamental insight into the particular character of the continuous manifolds of space and time:

"Definite parts of a manifold, which are distinguished from one another by a mark or boundary are called quanta. Their quantitative comparison is effected by means of counting in the case of discrete magnitudes and by measurement in the case of continuous ones.² Measurement consists in bringing the magnitudes to be compared into coincidence; for measurement, one therefore needs a means which can be applied (transported) as a standard of magnitude. If it is lacking, then two magnitudes can be compared only if one is a [proper] part of the other and then only according to more or less, not with respect to how much. . . . in the case of a discrete manifold, the principle [criterion] of the metric relations is already implicit in [intrinsic to] the concept of this manifold, whereas in the case of a continuous manifold, it must be brought in from elsewhere [extrinsically]. Thus, either the reality underlying space must form a discrete manifold or the reason for the metric relations must be sought extrinsically in binding forces which act on the manifold."

Russell [28, pp. 66–67] and the writer [13] have noted that, contrary to Riemann's apparent expectation, the first part of this statement will *not* bear critical scrutiny as a characterization of continuous manifolds *in general*. Riemann does, however, render here a fundamental feature of the continua of physical space and time, which are manifolds whose elements, taken singly, all have zero magnitude. And since our concern is with the geo-chronometry of continuous physical space and time, we can disregard defects in his account which do not affect its pertinence to the latter continua. By the same token, we can ignore inadequacies arising from his treatment of discrete and continuous types of order as *jointly exhaustive*. Instead, we state the valid upshot of his conception relevant to the spatio-temporal congruence issue before us. Construing his statement as applying, not only to lengths but also, *mutatis mutandis*, to areas and to volumes of higher dimensions, he gives the following

² Riemann apparently does not consider sets which are neither discrete nor continuous, but we shall consider the significance of that omission below.

sufficient condition for the intrinsic definability and non-definability of a metric *without* claiming it to be necessary as well: in the case of a discretely-ordered set, the "distance" between two elements can be defined *intrinsically* in a rather natural way by the cardinality of the "interval" determined by these elements.³ On the other hand, upon confronting the extended continuous manifolds of physical space and time, we see that neither the cardinality of intervals nor any of their other topological properties provide a basis for an *intrinsically*-defined metric. The first part of this conclusion was tellingly emphasized by Cantor's proof of the equi-cardinality of all positive intervals independently of their length. Thus, there is no *intrinsic* attribute of the space between the end-points of a line-segment AB , or any relation between these two points themselves, in virtue of which the interval AB could be said to contain the same amount of space as the space between the termini of another interval CD not coinciding with AB . Corresponding remarks apply to the time continuum. Accordingly, the continuity we postulate for physical space and time furnishes a *sufficient* condition for their *intrinsic metrical amorphousness*.⁴

³ The *basis* for the discrete ordering is not here at issue: it can be conventional, as in the case of the letters of the alphabet, or it may arise from special properties and relations characterizing the objects possessing the specified order.

⁴ Clearly, this does *not* preclude the existence of sufficient conditions *other than continuity* for the intrinsic metrical amorphousness of sets. But one cannot invoke densely-ordered, denumerable sets of points (instants) in an endeavor to show that discontinuous sets of such elements may likewise lack an intrinsic metric: even without measure theory, ordinary analytic geometry allows the deduction that the length of a *denumerably* infinite point set is intrinsically zero. This result is evident from the fact that since each point (more accurately, each unit point set or degenerate subinterval) has length *zero*, we obtain zero as the *intrinsic* length of the densely-ordered denumerable point set upon summing, in accord with the usual limit definition, the sequence of zero lengths obtainable by denumeration (cf. Grünbaum [11, pp. 297–298]). More generally, the measure of a denumerable point set is always zero (cf. Hobson [15, p. 166]) unless one succeeds in developing a very restrictive intuitionistic measure theory of some sort.

These considerations show incidentally that space-intervals cannot be held to be merely denumerable aggregates. Hence in the context of our post-Cantorean meaning of "continuous", it is actually not as damaging to Riemann's statement as it might seem *prima facie* that he neglected the *denumerable* dense sets by incorrectly treating the discrete and continuous types of order as *jointly exhaustive*. Moreover, since the distinction between denumerable and super-denumerable dense sets was almost certainly unknown to Riemann, it is likely that by "continuous" he merely intended the property which we now call "dense". Evidence of such an earlier usage of "continuous" is found as late as 1914: cf. Russell [27, p. 138].

The axioms of congruence [35, pp. 42–50] preempt “congruent” to be a spatial equality predicate but allow an infinitude of mutually-exclusive congruence classes of intervals. There are no *intrinsic* metric attributes of intervals, however, which could be invoked to single out *one* of these congruence classes as unique. Hence only the *choice* of a particular *extrinsic* congruence standard can determine a unique congruence class, the rigidity of that standard under transport being *decreed by convention*. And thus the role of this standard cannot be construed with Russell to be the mere ascertainment of an otherwise intrinsic equality obtaining between the intervals belonging to the congruence class defined by it. Similarly for time intervals and the periodic devices which define temporal congruence. And hence there can be no question at all of an *empirically* or factually determinate metric geometry or chronometry until *after* a physical stipulation of congruence.⁵

A concluding remark on the special importance of the equality term “congruent” (for line segments) vis-à-vis the other primitives of the calculus will precede turning our attention to some of the import of the conventionality of congruence.

Suitable alternative semantical interpretations of the term “congruent”, and correlatively of “straight line,” can readily demonstrate that, subject to the restrictions imposed by the existing topology, it is always a live option to give *either* a Euclidean *or* a *non-Euclidean* description of the same body of physico-geometrical facts. The possibility of alternative semantical interpretations of such *other* primitives of rival geometrical calculi as “point” does *not* generally have such relevance to this demonstration. Accordingly, when one is concerned, as we are here, with noting that, even apart from the logic of induction, the empirical facts themselves do *not* uniquely dictate the truth of either Euclidean geometry or of one of its non-Euclidean rivals, then the situation is as follows: the different physical interpretations of the term “congruent” (and hence of “straight line”) in the respective geometrical calculi enjoy a more central importance in the discussion than the semantics of such other primitives of these calculi as “point,” since the latter generally have the *same* physical meaning in both the Euclidean and non-Euclidean descriptions. Moreover, once we cease to look at physical geometry as a descriptively-interpreted system of abstract *synthetic* geometry and regard it instead as an interpreted system of abstract *differential* geometry of the

⁵ For a detailed critique of A. N. Whitehead’s *perceptualistic* objections to this conclusion [34, ch. VI; 35, ch. III; 36, *passim*] see Grünbaum [13].

Gauss-Riemann type, the pre-eminent status of the interpretation of "congruent" is seen to be beyond dispute: by choosing a particular distance function $ds = \sqrt{g_{ik}dx^i dx^k}$ for the line element, we specify not only what segments are congruent and what lines are straights (geodesics) but the entire geometry, since the metric tensor g_{ik} fully determines the Gaussian curvature K . To be sure, if one were discussing *not* the alternative between a Euclidean and non-Euclidean description of the same *spatial* facts but rather the set of *all* models (including *non*-spatial ones) of a *given* calculus, say the Euclidean one, then indeed the physical interpretation of "congruent" and of "straight line" would not merit any more attention than that of other primitives like "point".

The Import of Riemann's Conception of Congruence.

(a) F. Klein's Relative Consistency Proof of Hyperbolic Geometry and H. Poincaré's *Anschaulichkeitsbeweis* of that geometry.

In the light of the conventionality of congruence, F. Klein's relative consistency proof of hyperbolic geometry via a model furnished by the interior of a circle on the Euclidean plane ⁶ appears as merely one particular kind of possible remetrization of the circular portion of that plane, projective geometry having played the heuristic role of furnishing Klein with a suitable definition of congruence. What from the point of view of synthetic geometry appears as intertranslatability via a dictionary, appears as alternative metrizability from the point of view of differential geometry. Again, Poincaré's kind of *Anschaulichkeitsbeweis* of a three-dimensional hyperbolic geometry via a model furnished by the interior of a sphere in Euclidean space [20, pp. 75-8] is another example of remetrization. Here the alteration in our customary definition of congruence is conveyed to us pictorially by the effects of an inhomogeneous force field which appropriately shrinks all bodies alike as seen from the point of view of the normally Euclideanly-behaving bodies.

(b) Poincaré and the Conventionality of Congruence.

The central theme of Poincaré's so called conventionalism is essentially an elaboration of the thesis of alternative metrizability whose fundamental justification we owe to Riemann, and *not* [12, § 5] the *radical* conventionalism attributed to him by Reichenbach [23, p. 36].

Poincaré's much-cited and often misunderstood statement concerning the possibility of always giving a Euclidean description of any results of stellar parallax measurements is a less lucid statement of exactly the same point

⁶ For details, cf. Bonola [1, pp. 164-175]. For a summary of E. Beltrami's different relative consistency proof, see Struik [31, pp. 152-3].

made by him with magisterial clarity in the following passage [20, p. 235]:

"In space we know rectilinear triangles the sum of whose angles is equal to two right angles; but equally we know curvilinear triangles the sum of whose angles is less than two right angles. . . . To give the name of straights to the sides of the first is to adopt Euclidean geometry; to give the name of straights to the sides of the latter is to adopt the non-Euclidean geometry. So that to ask what geometry it is proper to adopt is to ask, to what line is it proper to give the name straight? It is evident that experiment can not settle such a question."

Now, the equivalence of this contention to Riemann's view of congruence becomes evident the moment we note that the legitimacy of identifying lines which are curvilinear in the usual geometrical parlance as "straights" is vouchsafed by the warrant for our choosing a new definition of congruence such that the previously curvilinear lines become geodesics of the new congruence. Corresponding remarks apply to Poincaré's contention that we can always preserve Euclidean geometry in the face of any data obtained from stellar parallax measurements: if the paths of light rays are geodesics on a particular definition of congruence, as indeed they are in the Schwarzschild procedure cited by Robertson, and if the paths of light rays are found parallaxically to sustain non-Euclidean relations on that metrization, then we need only choose a different definition of congruence such that these *same* paths will no longer be geodesics and that the geodesics of the newly chosen congruence are Euclideanly related. From the standpoint of synthetic geometry, the latter choice effects a *renaming* of optical and other paths and thus is merely a *recasting of the same factual content in Euclidean language rather than a revision of the extra-linguistic content of optical and other laws*⁷. Since Poincaré's claim here is a straightforward elaboration of the metric amorphousness of the continuous manifold of space, it is not clear how Robertson can reject it as a "pontifical pronouncement" and even regard it as being in contrast with what he calls Schwarzschild's "sound operational approach to the problem of physical geometry." [25, pp. 324-5]. For Schwarzschild had rendered the question concerning the prevailing geometry *factual* only by the adoption of a particular spatial

⁷ The remetrizability of Euclideanism affirmed by Poincaré [20, pp. 81-86] thus involves a *merely linguistic* interdependence of the geometric theory of rigid solids and the optical theory of light rays. This interdependence is logically *different*, as we shall see in Section 3, from P. Duhem's conception [6, Part II, ch. VI] of an *epistemological* interdependence, which Einstein espouses.

metrization based on the travel times of light, which does indeed turn the direct light paths of his astronomical triangle into geodesics.

There are two respects, however, in which Poincaré is open to criticism in this connection:

(i) He maintained [20, p. 81] that it would always be regarded as most convenient to preserve Euclidean geometry, even at the price of re-metrization, on the grounds that this geometry is the simplest analytically [20, p. 65]. Precisely the opposite development materialized in the general theory of relativity: Einstein forsook the simplicity of the geometry itself in the interests of being able to maximize the simplicity of the definition of congruence. He makes clear in his fundamental paper of 1916 that had he insisted on the retention of Euclidean geometry in a gravitational field, then he could *not* have taken "one and the same rod, independently of its place and orientation, as a realization of the same interval." [8, p. 161]

(ii) Even if the simplicity of the geometry itself were the sole determinant of its adoption, that simplicity might be judged by criteria other than Poincaré's analytical simplicity. Thus, Menger has urged that from the point of view of a criterion grounded on the simplicity of the undefined concepts used, hyperbolic and not Euclidean geometry is the simplest [16, p. 66].

On the other hand, if Poincaré were alive today, he could point to an interesting recent illustration of the sacrifice of the simplicity and accessibility of the congruence standard on the altar of maximum simplicity of the resulting theory. Astronomers have recently proposed to re-metrize the time continuum for the following reason: when the mean solar second, which is a very precisely known fraction of the period of the earth's rotation on its axis, is used as a standard of temporal congruence, then there are three kinds of discrepancies between the actual observational findings and those predicted by the usual theory of celestial mechanics. The empirical facts thus present astronomers with the following choice: Either they retain the rather natural standard of temporal congruence at the cost of having to bring the principles of celestial mechanics into conformity with observed fact by revising them appropriately. Or they re-metrize the time continuum, employing a less simple definition of congruence so as to preserve these principles intact. Decisions taken by astronomers in the last few years were exactly the reverse of Einstein's choice of 1916 as between the simplicity of the standard of congruence and that of the resulting theory. The mean solar second is to

be supplanted by a unit to which it is non-linearly related: the sidereal year, which is the period of the earth's revolution around the sun, due account being taken of the irregularities produced by the gravitational influence of the other planets.⁸

We see that the implementation of the requirement of descriptive simplicity in theory-construction can take alternative forms, because agreement of astronomical theory with the evidence now available is achievable by revising *either* the definition of temporal congruence *or* the postulates of celestial mechanics. The existence of this alternative likewise illustrates that for an axiomatized physical theory containing a geochronometry, it is *gratuitous* to single out the postulates of the theory as having been prompted by *empirical* findings in contradistinction to deeming the *definitions of congruence* to be wholly *a priori*, or vice versa. This conclusion bears out geochronometrically Braithwaite's contention in this volume that there is an important sense in which axiomatized physical theory does not lend itself to compliance with Heinrich Hertz's injunction to "distinguish thoroughly and sharply between the elements . . . which arise from the necessities of thought, from experience, and from arbitrary choice." [14, p. 8].⁹

(c) The impossibility of defining congruence uniquely by stipulating a particular metric geometry.

A question which arises naturally upon undertaking the mathematical implementation of a given choice of a metric geometry in the context of a particular set of topological facts is the following: do these facts in conjunction with the desired metric geometry determine a unique definition of congruence? If the answer were actually in the affirmative, as both Carnap [3, pp. 54–55] and Reichenbach [23, pp. 33–34; 22, pp. 132–133] have maintained, this would mean that the desired geometry would uniquely specify a metric tensor under given factual circumstances and thus, in a particular coordinate system, a unique set of functions g_{ik} . But Carnap's and Reichenbach's assertion of uniqueness is erroneous, as is demonstrated by showing that besides the customary definition of congruence, which assigns the same length to the measuring rod everywhere and thereby confers a Euclidean geometry on an ordinary table top, there are infinitely many *other* definitions of congruence which likewise

⁸ For a clear account of the relevant astronomical details, see Clemence [4].

⁹ Braithwaite's point was made independently by Pap [19], who argues that the analytic-synthetic distinction cannot be upheld for partially-interpreted theoretical languages like that of theoretical physics.

yield a Euclidean geometry for that surface but which make the length of a rod depend on its orientation or position. Thus, consider our horizontal table top equipped with a net-work of Cartesian coordinates x and y and suppose that another such surface intersects the horizontal one at an angle θ so that their line of intersection is both the y -axis of the horizontal plane and the \bar{y} -axis of a rectangular system of coordinates \bar{x} and \bar{y} on the inclined plane. Assume that the inclined plane has been metrized in the customary way. But then remetrize the *horizontal* plane by calling congruent in it those line segments which are the perpendicular projections onto it of segments of the inclined plane that are equal in the latter's metric. Accordingly, we have a mapping

$$\bar{x} = x \sec \theta$$

$$\bar{y} = y,$$

and we now assign to a line segment of the horizontal plane whose termini have the coordinate differences dx and dy *not* the customary length $\sqrt{dx^2 + dy^2}$ but rather

$$ds = \sqrt{d\bar{x}^2 + d\bar{y}^2} = \sqrt{\sec^2 \theta dx^2 + dy^2}.$$

Nonetheless, upon using the *new* g_{ik} , which are introduced into the x, y coordinates by the revised definition of congruence, to compute the Gaussian curvature of the horizontal table top, we still obtain the Euclidean value zero. And by merely varying the angle of inclination θ , we obtain infinitely many *different* definitions of congruence all of which make the length of a given rod dependent on its orientation and yet impart a Euclidean geometry to the horizontal table top. Thus, the requirement of Euclideanism does not uniquely determine a metric tensor, and, contrary to Carnap and Reichenbach, *there are infinitely many ways in which a measuring rod could squirm under transport as compared to its customary behavior and still yield a Euclidean geometry*. In fact, even for plane Euclidean geometry, the class of congruence definitions is far wider than the one-parameter family yielded by our particular isometric mappings of an inclined plane onto the horizontal one. Dr. Samuel Gulden, to whom I presented the problem of determining the class of different metric tensors for *each* kind of two-dimensional and three-dimensional Riemannian space, has pointed out that (i) in the Euclidean case, upon abandoning the restriction of our above isometric mappings to affine coordinate transformations and considering non-linear transformations with non-vanishing Jacobian, we can generate infinitely many other

metrizations whose associated Gaussian curvature is everywhere zero. For example, for the admissible transformation between our two sets of rectangular coordinates x, y and \bar{x}, \bar{y} given by

$$\bar{x} = x + \frac{1}{3}y^3, \text{ and}$$

$$\bar{y} = \frac{1}{3}x^3 - y,$$

the distance function becomes

$$ds^2 = d\bar{x}^2 + d\bar{y}^2 = (1 + x^4)dx^2 + 2(y^2 - x^2)dxdy + (y^4 + 1)dy^2.$$

In this case, the length of a given rod is generally dependent both on its position and on its orientation, (ii) the result obtained for Euclidean space can be generalized to a very large class of Riemann spaces of various dimensions.

We are now ready to consider the *second* of the two principal problems which have been posed in connection with the criterion of rigidity.

3. The Criterion of Rigidity: II. The Logic of Correcting for "Distorting"

Influences. Physical geometry is usually conceived as the system of metric relations exhibited by transported solid bodies *independently* of their particular chemical composition. On this conception, the criterion of congruence can be furnished by a transported solid body for the purpose of determining the geometry by measurement, only if the computational application of suitable "corrections" (or, ideally, appropriate shielding) has essentially eliminated inhomogeneous thermal, elastic, electric and other influences, which produce changes of *varying degree* ("distortions") in different kinds of materials. The demand for this *elimination* as a prerequisite to the experimental determination of the geometry has a thermodynamic counterpart: the requirement of a means for measuring temperature which does not yield the discordant results produced by expansion thermometers at other than fixed points when different thermometric substances are employed. This thermometric need is fulfilled successfully by Kelvin's thermodynamic scale of temperature. But attention to the implementation of the corresponding prerequisite of physical geometry has led Einstein [9, pp. 676-678] to impugn the empirical status of that geometry. He considers the case in which congruence has been defined by the diverse kinds of transported solid measuring rods *as corrected for their respective idiosyncratic distortions* with a view to *then* making an empirical determination of the prevailing geometry. And in an

argument which he attributes to Poincaré, Einstein's thesis is that the very logic of computing these corrections precludes that the geometry itself be accessible to experimental ascertainment in isolation from *other* physical regularities. Specifically, he states the case in the form of a dialogue between Reichenbach and Poincaré¹⁰:

Poincaré: The empirically given bodies are not rigid, and consequently can not be used for the embodiment of geometric intervals. Therefore, the theorems of geometry are not verifiable.

Reichenbach: I admit that there are no bodies which can be *immediately* adduced for the "real definition" of the interval. Nevertheless, this real definition can be achieved by taking the thermal volume-dependence, elasticity, electro- and magneto-striction, etc., into consideration. That this is really [and] without contradiction possible, classical physics has surely demonstrated.

Poincaré: In gaining the real definition improved by yourself you have made use of physical laws, the formulation of which presupposes (in this case) Euclidean geometry. The verification, of which you have spoken, refers, therefore, not merely to geometry but to the entire system of physical laws which constitute its foundation. An examination of geometry by itself is consequently not thinkable. — Why should it consequently not be entirely up to me to choose geometry according to my own convenience (i.e., Euclidean) and to fit the remaining (in the usual sense "physical") laws to this choice in such manner that there can arise no contradiction of the whole with experience?"

The objection which Einstein presents here on behalf of conventionalism is aimed at a conception of physical geometry which is *empiricist* merely in Carnap's and Reichenbach's *conditional* sense explained in Section 1. Einstein's criticism is that the rigid body is not even defined without first *decreeing* the validity of Euclidean geometry. And the grounds he gives for this conclusion are that *before* the *corrected* rod can be used to make an *empirical* determination of the *de facto* geometry, the required corrections must be computed via laws, such as those of elasticity, which involve *Euclideanly*-calculated areas and volumes. But clearly the warrant

¹⁰ It is rather doubtful that Poincaré himself espoused the version of conventionalism which Einstein links to his name here: in speaking of the variations which solids exhibit under distorting influences, Poincaré says [20, p. 76]: "we neglect these variations in laying the foundations of geometry, because, besides their being very slight, they are irregular and consequently seem to us accidental."

for thus introducing Euclidean geometry *at this stage* cannot be empirical.

I now wish to set forth my reasons for believing that Einstein's argument does not succeed in making physical geometry a matter of convention rather than fact in a sense which is *independent* of the alternative metrizable vouchsafed by spatio-temporal continuity.

There is no question that the laws used to make the corrections for deformations [30, p. 60; 32, p. 408] involve areas and volumes in a fundamental way (e.g. in the definitions of the elastic stresses and strains) and that this involvement presupposes a geometry, as is evident from the area and volume formulae

$$A = \int \sqrt{g} \, dx^1 dx^2 \text{ and } V = \int \sqrt{g} \, dx^1 dx^2 dx^3,$$

where "g" represents the determinant of the components g_{ik} [10, p. 177]. Now suppose that we begin with a set of Euclidean-formulated physical laws P_0 in correcting for the distortions induced by perturbations and then use the thus Euclidean-corrected congruence standard for *empirically* exploring the geometry of space by determining the metric tensor. *The initial stipulational affirmation of the Euclidean geometry G_0 in the physical laws P_0 used to compute the corrections in no way assures that the geometry obtained by the corrected rods will be Euclidean!* If it is non-Euclidean, then the question is: what will Einstein's fitting of the physical laws to preserve Euclideanism and avoid a contradiction of the total theoretical system with experience involve? Will the adjustments in P_0 necessitated by the retention of Euclideanism entail merely a change in the dependence of the length assigned to the transported rod on such *non-positional* parameters as temperature, pressure, magnetic field etc.? Or could the putative empirical findings compel that the length of the transported rod be likewise made a function of its *position* and *orientation* in order to square the coincidence findings with the requirement of Euclideanism? The temporal variability of distorting influences and the possibility of obtaining non-Euclidean results by measurements carried out in a spatial region uniformly characterized by standard conditions of temperature, pressure, electric and magnetic field strength etc. show it to be quite doubtful that the preservation of Euclideanism could always be accomplished short of introducing the dependence of the rod's length on position and orientation. Thus, the need for *remetrizing* in this sense in order to retain Euclideanism cannot be ruled out. But this kind of remetrization does not provide the requisite support for Einstein's version of conventionalism, whose onus it is to show that the geometry by itself

cannot be held to be empirical even when we *exclude* resorting to such remetrization.

That the geometry may well be empirical in this sense is seen from the following possibilities of its successful empirical determination. After assumedly obtaining a non-Euclidean geometry G_1 from measurements with a rod corrected on the basis of Euclideanly-formulated physical laws P_0 , we can revise P_0 so as to conform to the non-Euclidean geometry G_1 just obtained by measurement. This retroactive revision of P_0 would be effected by recalculating such quantities as areas and volumes on the basis of G_1 and changing the functional dependencies relating them to temperature and other physical parameters. We thus obtain a new set of laws P_1 . Now we use this set P_1 of laws to correct the rods for perturbational influences and then determine the geometry with the thus corrected rods. If the result is a geometry G_2 different from G_1 , then *if there is convergence to a geometry of constant curvature*, we must repeat this process a finite number of times until the geometry G_n ingredient in the laws P_n providing the basis for perturbation-corrections is indeed the same to within experimental accuracy as the geometry obtained by measurements with rods that have been corrected via the set P_n .

If there is such convergence at all, it will be to the same geometry G_n even if the physical laws used in making the initial corrections are not the set P_0 , which presupposes Euclidean geometry, but a different set P based on some non-Euclidean geometry or other. That there can exist only one such geometry of constant curvature G_n would seem to be guaranteed by the identity of G_n with the unique underlying geometry G_t characterized by the following properties: (i) G_t would be exhibited by the coincidence behavior of a transported rod if the *whole* of the space were actually free of deforming influences, (ii) G_t would be obtained by measurements with rods corrected for distortions on the basis of physical laws P_t presupposing G_t , and (iii) G_t would be found to prevail in a given relatively small, perturbation-free region of the space quite independently of the assumed geometry ingredient in the correctional physical laws. Hence, *if our method of successive approximation does converge to a geometry G_n of constant curvature, then G_n would be this unique underlying geometry G_t .* And, in that event, we can claim to have found empirically that G_t is indeed the geometry prevailing in the entire space which we have explored.

But what if there is no convergence? It might happen that whereas convergence would obtain by starting out with corrections based on the

set P_0 of physical laws, it would *not* obtain by beginning instead with corrections presupposing some particular *non*-Euclidean set P or *vice versa*: just as in the case of Newton's method of successive approximation [5, p. 286], there are conditions, as A. Suna has pointed out to me, under which there would be no convergence. We might then nonetheless succeed as follows in finding the geometry G_t empirically, *if* our space is one of constant curvature.

The geometry G_r resulting from measurements by means of a corrected rod is a single-valued function of the geometry G_a assumed in the correctional physical laws, and a Laplacian demon having sufficient knowledge of the facts of the world would know this function $G_r = f(G_a)$. Accordingly, we can formulate the problem of determining the geometry empirically as the problem of finding the point of intersection between the curve representing this function and the straight line $G_r = G_a$. That there exists one and only one such point of intersection follows from the existence of the geometry G_t defined above, provided that our space is one of constant curvature. Thus, what is now needed is to make determinations of the G_r corresponding to a number of geometrically-different sets of correctional physical laws P_a , to draw the most reasonable curve $G_r = f(G_a)$ through this finite number of points (G_a, G_r) , and then to find the point of intersection of this curve and the straight line $G_r = G_a$.

Whether this point of intersection turns out to be the one representing Euclidean geometry or not is beyond the reach of our conventions, *barring* a remetrization. And thus the least that we can conclude is that since empirical findings can greatly narrow down the range of uncertainty as to the prevailing geometry, there is no assurance of the *latitude* for the choice of a geometry which Einstein takes for granted. Einstein's Duhemian position would appear to be inescapable *only* if our proposed method of determining the geometry by itself empirically *cannot* be generalized in some way to cover the general relativity case of a space of *variable* curvature and if the latter kind of theory turns out to be true.

It would seem therefore that, contrary to Einstein, the logic of eliminating distorting influences prior to stipulating the rigidity of a solid body is *not* such as to provide scope for the ingression of conventions over and above those acknowledged in Riemann's analysis of congruence, and trivial ones such as the system of units used. Our analysis of the logical status of the concept of a rigid body thus leads to the conclusion that once the physical meaning of congruence has been stipulated by reference to a solid body for whose distortions allowance has been made compu-

tationally as outlined, then the geometry is determined uniquely by the totality of relevant empirical facts. It is true, of course, that even apart from experimental errors, not to speak of quantum limitations on the accuracy with which the metric tensor of *space-time* can be meaningfully ascertained by measurement [29; 37], no *finite* number of data can uniquely determine the functions constituting the representations g_{ik} of the metric tensor in any given coordinate system. But the criterion of *inductive* simplicity which governs the free creativity of the geometer's imagination in his choice of a particular metric tensor here is the same as the one employed in theory formation in any of the non-geometrical portions of empirical science. And choices made on the basis of such inductive simplicity are in principle true or false, unlike those springing from considerations of descriptive simplicity, which merely reflect conventions.

The author is indebted to Dr. Samuel Gulden of the Department of Mathematics, Lehigh University, U.S.A. for very helpful discussions.

Bibliography

- [1] BONOLA, R., *Non-Euclidean Geometry*. New York, 1955. IX + 268 pp.
- [2] BROWN, F. A., *Biological clocks and the fiddler crab*. Scientific American, vol. 190 (April, 1954), pp. 34–37.
- [3] CARNAP, R., *Der Raum*. Berlin, 1922 (Supplement No. 56 of *Kant-Studien*) 87 pp.
- [4] CLEMENCE, G. M., *Time and its measurement*. American Scientist, vol. 40 (1952), pp. 260–269; and *Astronomical time*. Reviews of Modern Physics, vol. 29 (1957), p. 5.
- [5] COURANT, R., *Vorlesungen über Differential- und Integralrechnung*, vol. 1. Berlin, 1927. XIV + 410 pp.
- [6] DUHEM, P., *The Aim and Structure of Physical Theory*. Princeton, 1954. XXII + 344 pp.
- [7] EDDINGTON, A. S., *Space, Time and Gravitation*. Cambridge, 1953. VII + 218 pp.
- [8] EINSTEIN, A., *The foundations of the general theory of relativity*. In: *The Principle of Relativity*, a collection of original memoirs, London, 1923, pp. 111–164.
- [9] —, *Reply to criticisms*. In: *Albert Einstein: Philosopher-Scientist* (edited by SCHILPP, P. A.) Evanston, 1949, pp. 665–688.
- [10] EISENHART, L. P., *Riemannian Geometry*. Princeton, 1949. VII + 306 pp.
- [11] GRÜNBAUM, A., *A consistent conception of the extended linear continuum as an aggregate of unextended elements*. Philosophy of Science, vol. 19 (1952), pp. 288–306.

- [12] —, *Carnap's views on the foundations of geometry*. In: *The Philosophy of Rudolf Carnap* (edited by SCHILPP, P. A.), (forthcoming).
- [13] —, *Geometry, Chronometry and Empiricism*. In: *Minnesota Studies in the Philosophy of Science* (edited by FEIGL, H. and MAXWELL, G.), vol. III (forthcoming).
- [14] HERTZ, H., *The Principles of Mechanics*. New York, 1956, 271 pp.
- [15] HOBSON, E. W., *The Theory of Functions of a Real Variable*, vol. 1. New York, 1957, XV + 736 pp.
- [16] MENGER, K., *On algebra of geometry and recent progress in non-euclidean geometry*. The Rice Institute Pamphlet, vol. 27 (1940), pp. 41–79.
- [17] MILNE, E. A., *Kinematic Relativity*. Oxford, 1948, VI + 238 pp.
- [18] NEWTON, I., *Principia* (edited by CAJORI, F.). Berkeley, 1947, XXXV + 680 pp.
- [19] PAP, A., *Are physical magnitudes operationally definable?* In: *Measurement: Definitions and Theories* (edited by CHURCHMAN, C. W. and RATOOSH, P.) New York, 1959 (in press).
- [20] POINCARÉ, H., *The Foundations of Science*. Lancaster 1946, XI + 553 pp.
- [21] REICHENBACH, H., *The philosophical significance of the theory of relativity*. In: *Albert-Einstein: Philosopher-Scientist* (edited by SCHILPP, P. A.) Evanston, 1949, pp. 287–311.
- [22] —, *The Rise of Scientific Philosophy*. Berkeley, 1951, XI + 333 pp.
- [23] —, *The Philosophy of Space and Time*. New York, 1958, XVI + 295 pp.
- [24] RIEMANN, B., *Gesammelte Mathematische Werke* (edited by WEBER and DEDEKIND). New York, 1953, X + 558 pp.
- [25] ROBERTSON, H. P., *Geometry as a branch of physics*. In: *Albert Einstein: Philosopher-Scientist* (edited by SCHILPP, P. A.). Evanston, 1949, pp. 313–332.
- [26] RUSSELL, B., *Sur les axiomes de la géométrie*. Revue de Métaphysique et de Morale, vol. 7 (1899), pp. 684–707.
- [27] —, *Our Knowledge of the External World*. London, 1926, 251 pp.
- [28] —, *The Foundations of Geometry*. New York, 1956, 201 pp.
- [29] SALECKER, H. and WIGNER, E. P., *Quantum Limitations of the Measurement of Space-Time Distances*. The Physical Review, vol. 109 (1958), pp. 571–577.
- [30] SOKOLNIKOFF, I. S., *Mathematical Theory of Elasticity*. New York, 1946, XI + 373 pp.
- [31] STRUIK, D. J., *Classical Differential Geometry*. Cambridge, 1950, VIII + 221 pp.
- [32] TIMOSHENKO, S. and GOODIER, J. N., *Theory of Elasticity*. New York, 1951, XVIII + 506 pp.
- [33] VON HELMHOLTZ, H., *Schriften zur Erkenntnistheorie* (edited by HERTZ, P. and SCHLICK, M.). Berlin, 1921, IX + 175 pp.
- [34] WHITEHEAD, A. N., *The Concept of Nature*. Cambridge, 1926, VIII + 202 pp.
- [35] —, *The Principle of Relativity*. Cambridge, 1922, XII + 190 pp.
- [36] —, *Process and Reality*. New York, 1929, XII + 546 pp.
- [37] WIGNER, E. P., *Relativistic invariance and quantum phenomena*. Reviews of Modern Physics, vol. 29 (1957), pp. 255–268.

PART II
FOUNDATIONS OF PHYSICS

HOW MUCH RIGOR IS POSSIBLE IN PHYSICS?

P. W. BRIDGMAN

Harvard University, Cambridge, Massachusetts, U.S.A.

Let me begin by saying that I have accepted the invitation to speak to this Symposium on the Axiomatic Method with extreme hesitation. I think I realize that there is a highly developed axiomatic technique and that to many of you the questions of greatest interest in this field are questions of technique. To an outsider like myself the spectacle of the virtuosity exhibited by some of you in the practise of this technique is a little terifying. I realize that many, if not all of you, will be impatient with the generalities which I have to offer and will be eager to get on with the more vital business of detailed attack on the numerous technical problems. I cannot even hope that my generalities will not seem to you too obvious to be worth saying, and that I may appear in the light of an enfant terrible, blurting out the things that everyone knows but has too much sense to say out loud. If, in spite of all this, I am venturing to talk to you, it is partly selfish because it appeared that I could not otherwise attend this meeting, and I expect, in spite of your technicalities, to pick up points of view which will be new and profitable. But beyond this, I do think that it is worth while, occasionally, to say the obvious things out loud, for I do not believe that we have, even yet, taken into account all the obvious things. In any event, I am glad that the program committee put my paper in the opening session, so that you can soon get it out of the way and turn to more interesting and pressing matters.

The "rigor" which I shall talk about is not itself a very precise or rigorous thing. In its first usage "rigor" is applied to reasoning. If, however, rigorous reasoning is to be possible, the objects and operations of our reasoning must have certain properties, so that "rigor" comes to have an extended meaning. In this extended meaning it implies sharpness and precision and it has overtones of certainty. It is in this extended sense that I shall be concerned with rigor. My task will be to examine to what extent what we do in physics can have the attributes of sharpness, precision, and certainty. I shall assume as not needing argument that in no field of activity are these attributes actually attainable, but they

function only as limiting ideals, which are never fully attained even in as abstract a domain as that of postulate theory.

All human enterprise, of which postulate theory and physics are special cases, is subject to one restriction on any attainable sharpness or certainty which is so ubiquitous and unavoidable that we seldom bother even to mention it. The possibility of self-doubt is always with us; we can always ask ourselves whether we are really doing what we think we are doing or how we can be sure that we have not suddenly gone insane or are not dreaming. All our intellectual activity not only is, but has to be, based on the premise that intellectually we are going concerns. In so far as this is common to postulate theory and the physics of the laboratory I need not stop to elaborate the point further. It seems to me, however, that there are points here which in another context might be analyzed further than they usually are. Just what is involved in the assumption that I am a going concern intellectually? and how shall I go to work to assure myself that the assumption actually applies to me? In particular what is the method by which I can assure myself that I am not now dreaming? I have seen no such method.

Forgetting now any lack of sharpness arising from self doubt, there are certain human activities which apparently have perfect sharpness. The realm of mathematics and of logic is such a realm, par excellence. Here we have yes-no sharpness — two numbers are either equal to each other or they are not; a certain point either lies on a given line or it does not; there is only *one* straight line connecting any two points. Now it is a matter of observation that this yes-no sharpness is found only in the realm of things we say as distinguished from the realm of things we do. Sharpness is an attribute of the way we talk about our experience, in particular whether we talk about it in yes-no terms, rather than an attribute of the experience itself, if you will be charitable enough to grant me meaning in such a way of expression. Nothing that happens in the laboratory corresponds to the statement that a given point is either on a given line or it is not.

There is no question but that we do talk about aspects of experience in yes-no terms, and in so far as any field of experience has such yes-no sharpness it has it in virtue of the fact that it is a verbal activity. One may well question, however, whether we have any right to ascribe such yes-no properties to any verbal activity. What are these words anyhow? They are not static things, but are themselves a form of activity which varies in some way with every so-called repetition of the word. A word as we use

it is part of a terribly complicated system, involving both present structure in the brain and the past experience of the brain, most of which we cannot possibly be conscious of. The assumption that we are going concerns intellectually involves much more than merely the absence of self doubt.

The physics of measurement and of the laboratory does not have the yes-no sharpness of mathematics, but nevertheless employs conventional mathematics as an indispensable tool. Every physicist combines in his own person, to greater or less degree, the experimental physicist who makes measurements in the laboratory, and the theoretical physicist who represents the results of the measurements by the numbers of mathematics. These numbers are things that he says or writes on paper. The jump by which he passes from the operations of the laboratory to what he mathematically says about the operations is a jump which may not be bridged logically, and is furthermore a jump which ignores certain essential features of the physical situation. For the mathematics which the physicist uses does not exactly correspond to what happens to him. In the laboratory every measurement is fuzzy because of error. As far as reproducing what happens to him is concerned, the mathematics of the physicist might equally well be the mathematics of the rational numbers, in which such irrationals as $\sqrt{2}$ or π do not occur. Now one would certainly be going out of one's way to attempt to force theoretical physics into a straightjacket of the mathematics of the rational numbers as distinguished from the mathematics of all real numbers, but by forcing it into the straight jacket of any kind of mathematics at all, with its yes-no sharpness, one is discarding an essential aspect of all physical experience and to that extent renouncing the possibility of exactly reproducing that experience. In this sense, the commitment of physics to the use of mathematics itself constitutes, paradoxically, a renunciation of the possibility of rigor.

The unavoidable presence of error in any physical measurement which we are here insisting on reminds one of the fuzziness in the measurement of conjugate quantities covered by the Heisenberg principle of indetermination, but is, I believe, something quite different. The sort of error that we here are concerned with would still be present in our knowledge of the so-called "pure case" of quantum mechanics. In so far as quantum theory treats, for example, the charge on the electron or Planck's constant as mathematically sharp numbers, as it does, it is in so far neglecting an essential aspect of all our experience. It used to be thought that the errors

of physical measurement were a more or less irrelevant epiphenomenon, which could be avoided in the limit by the construction of better and better measuring apparatus. This happy conviction appeared less compelling when the atomic structure of all matter was established, including the atomic structure of the measuring apparatus. Now, it appears to me, the linkage of error with every sort of physical measurement must be regarded as inevitable when it is considered that the knowledge of the measurement, which is all we can be concerned with, is a result of the coupling of the external situation with a human brain. Even if we had adequate knowledge of the details of this coupling we admittedly could not yet use this knowledge in formulating in detail how the unavoidable fuzziness should be incorporated in our description of the world nor how we should modify our present use of mathematics. About the only thing we can do at present is to continue in our present use of mathematics, but with the addition of a *caveat* to every equation, warning that things are not quite as they seem.

Quantum theory has effectively called to attention certain other important features of the world about us. The realm in which quantum effects are usually considered to be important is in the first instance the realm of small things — small distances and short times. Phenomena in this realm do not present themselves directly to our unaided senses, but occur only in conjunction with special types of instrument, with which we say that we “extend” the scope of our senses. But if we examine what we actually do, we see that these instruments function through our conventional senses. Hence, it does not reproduce what actually happens to say, for example, that the microscope reveals to us a new “microscopic world”. The so-called microscopic world is really a new macroscopic world which we have found how to enter by inventing new kinds of macroscopic instrument. The “world” of quantum phenomena eventually has to find its description and explanation in terms of the things that happen to us on the macroscopic scale of every day life. I think most quantum theorists will admit this if they are pressed, but in spite of this the language of ordinary quantum theory is a language of microscopic entities which we handle verbally just as if they had the existential status of the objects of daily life. There is ample justification for this in the enormous simplification which results in our description and our handling of experience. This simplification is nevertheless bought at a price — the price of neglecting and forgetting some of the unavoidable accompaniments of all our experience. By thus agreeing to blur some of the recog-

nizable aspects of experience we have at the same time condemned ourselves to a loss of possible rigor, using "rigor" with the implications already explained. This sort of thing is by no means characteristic exclusively of quantum theory — strictly we should never think of bacteria without thinking of microscopes or think of galaxies without thinking of telescopes, but such rigor of thought is hardly attainable in practise.

Another matter which quantum theory has forcibly called to our attention is that the instrument of observation may not properly be separated from the object of observation. Heisenberg's principle of indetermination is one of the consequences of following out the implications of this. The principle that instrument of observation is not to be separated from object of observation is, it seems to me, a special case of a broader principle, namely that experience has to be taken as a whole and may not be analyzed into pieces. In other words, the operation of isolation is not a legitimate operation. Now the operation of isolation is perhaps the most universal of all intellectual operations, and without it rational thought would hardly be possible. Nevertheless, in the world of quantum phenomena situations arise in which our propensity for isolating definitely gets us into trouble. For instance, the electron is not properly to be thought of in isolation, but only as an aspect of the total experimental set-up in which it appears. When we view the electron in this light the paradox disappears from such situations as the interference pattern formed by the electron in the presence of two slits, where the electron, if we treat it as an ordinary isolatable object that can go through only one of the slits, apparently "knows" of the existence of the other slit without going through it. We are thus driven to concede that the operation of isolation cannot be legitimate "in principle", but this concession presents us with an extraordinarily difficult dilemma, for the very words in which we express the illegitimacy of the operation of isolation receive their meaning only in a context of isolation. In practise we meet the situation as best we can by methods largely intuitive in character which we have acquired by long practise. But I think that even our best practise has disclosed no method of sharply handling the situation — the method of isolation is neither sharply separated from the method of holism, nor is there any sharp criterion which determines when we shall shift from the one method to the other. Neither is it possible to express sharply in language what we mean by the one as distinguished from the other. The best we can do in practise is a sort of spiralling approximation, shifting back and forth from one level of operation to the other, and concentrating our attention

first on one aspect and then on another of the total situation. In such a setting we cannot expect rigor.

There are many other situations in which the operation of isolation leads to dilemma and paradox. Long ago, on the classical level, the concepts of thermodynamics found their meaning in terms of operations performed on isolated systems. Not only do the fundamental concepts of energy and entropy receive their meaning in terms of physical systems isolated in space, but isolation in time is also required, because otherwise reversibility, or, more generally, recoverability of previous condition, does not occur. Without recoverability the concepts of thermodynamics are incapable of definition. This necessity for isolation in the fundamental definitions leads to logical difficulties when we attempt to extend the notions of energy or entropy to the universe as a whole. The logical status of any theorem involving the conservation of the energy of the universe, or the universal degradation of energy and eventual heat death of the universe, seems to me exceedingly obscure. Furthermore, the classical connection between deterministic and statistical mechanics which expresses entropy in probabilistic terms seems to me to involve an illegitimate treatment of the entropy of isolated bodies. It is often said that an isolated system comprising many molecules approaches, with the passage of time, a completely disordered state and hence the condition of maximum entropy, because of the "law of large numbers", in virtue of which the internal condition eventually becomes one of molecular chaos in spite of the fact that the laws of the individual molecular encounters are completely deterministic. This it seems to me is logically fallacious. Given an isolated system, with a definite initial distribution and deterministic individual encounters, logically it can never evolve into a system with chaotic distribution. To say that chaos gets in through the operation of the "law of large numbers" seems to me to introduce a completely unjustified and ad hoc concept. But chaos *may* logically get into the system through the walls which are coupled to the external world. This coupling is part of a divergent process — the state of the walls may not be deterministically specified except by coupling them to an ever increasing domain of the external world over which we have ever less control. The only acceptable method which has been found for dealing with this divergent process is through probability. Here again we have paradox — the concept, entropy, is applicable only in a context of isolated systems, but the detailed mechanism, through the operation of which entropy functions, occurs only in non-isolated systems.

In general it seems to me that the situations contemplated in probability analysis are particularly situations in which the jump from theory to application cannot be made sharply, so that no application of probability theory can be rigorous. It is particularly important to realize this now that quantum theory is disposed to regard probabilities as something fundamental and unanalyzable rather than as an artefact in an essentially deterministic universe. Against this tendency of the theoretical physicists must be placed, I believe, the realization that the fundamental concepts of probability have meaning only in the context of a deterministic background. No situation is ever *completely* chaotic, but it is only restricted aspects which are probabilistic. We cannot say that a particular fall of a die is undetermined and probabilistic unless the die itself, the table top on which it rolls, and we ourselves who observe it and talk about it, retain their conventional deterministic identity. We have here a special case of the theorem that eventually any new concepts must find their meaning on the level of daily life. And since the level of daily life is preponderantly deterministic, I believe it is impossible to handle probability consistently as ultimate and unanalyzable.

"Randomness" is a concept fundamental in probability analysis and of such importance that lists of random numbers are often printed and employed in practical applications. Yet theoretically no finite set of numbers can be completely random, because there are an infinite number of conditions of randomness. In practise, no set of numbers that has been printed or otherwise actually exhibited can possibly be random, nor can a series of events that has actually occurred be random, because it is always possible to find some sort of regularity in any finite sequence. The concept of randomness, so fundamental to the whole conceptual edifice, thus appears as a loose concept, incapable of realization in practise. "Randomness" occurs only in the realm of things we say.

Probability theory runs into other sorts of difficulty when it deals with rare events. A literal application of kinetic theory and statistical mechanics yields a calculably small finite probability for any compound event. Thus there is a finite probability that if we watch long enough we shall some day see a pail of water freeze on the fire, a conclusion that Bertrand Russell has delighted to rub in. Or consider another example in somewhat the same vein. Suppose that I have measured some object by ordinary laboratory procedures and find it to be 1.500 meters, with some apparent uncertainty in the last millimeter. Suppose that I choose to report this measurement by saying that the length of the object is between 1 and 2

meters. Then probability theory states that there is *some* probability that this statement is incorrect. Now it seems to me that a theory which makes these two statements, about the freezing water and the error of my measurement, is a theory which fails to agree *qualitatively* with the nature of everyday experience. The finite probability of freezing or of error is a property of our mathematics, not of the situation which the mathematics is designed to describe, and in thus dealing with rare events our probability analysis reveals itself as only an approximation. In general, it seems to me that one has a right to question any probability analysis which predicts an event so rare that it has not yet been observed. One might even venture a theorem to this effect. Such a putative theorem receives a certain justification when it is considered that the prediction of rare events involves long range extrapolations, which would demand the establishment of the fundamental laws of mechanics with an accuracy far beyond that actually attainable.

Our intellectual difficulties are thus not peculiar to the new situations revealed by quantum theory, but classical physics has always had its share of difficulty and paradox. Among these difficulties may be mentioned these of dealing with continuous media. The equations of hydrodynamics, for instance, purportedly deal with continuous media, but the variables in the equations refer to the motion of "particles" of the fluid, which, whatever other properties they may have, at least have the property of identifiability. Whatever it is that bestows the identifiability would seem to violate the presumptive perfect homogeneity and continuity of the fluid. The two concepts are mutually contradictory and exclusive, but nevertheless our thinking seems to demand them, and as far as I know no one has invented a way of getting along without them.

I believe that there are somewhat similar difficulties with the concept of "field" which by many is regarded as fundamental to modern theoretical physics. We think of the field at any point of space as something "real", independently of whether there is an instrument at the point to measure it. But when we try to account mathematically for the fact that our instrument apparently responds to what was there before we went there with the instrument, we find that actually the instrument responds to the modified state of affairs after the instrument is introduced. (This is shown by an analysis of the Maxwell stresses.) Our attempt to give instrumental meaning to something that exists in the absence of the instrument seems foredoomed to failure — one can detect the odor¹ of a logical inconsistency here. Yet our thinking seems to demand that we

attach a meaning to what would be there in the absence of the instrument, whereas meaning itself exists only in a context of instruments.

All the infelicities and ineptnesses which we have encountered up to now arise because we have been trying to do something with our minds which cannot be done. After long experience we have found how to deal with situations of this sort after a fashion. We push the conventional line of attack as far as we can, and when we presently run into conceptual difficulties, we usually meet these difficulties, not by any drastic revision of our conceptual structure, but by keeping as much of it as we can and patching it up by rules explicitly warning of the limitations of the conventional machinery. There is a certain resemblance between this general situation and the special situations in quantum theory to which the Heisenberg principle is applicable. We cannot, for example, push our conventional description of a physical system in terms of space and time too far toward the microscopic without running into difficulties with our description of the same system in terms of cause and effect, although on the scale of daily life a description in terms of space and time is practically synonymous with a description in terms of cause and effect. There are many examples in quantum theory where we have to decide between which of two mutually exclusive forms of description we shall employ. Bohr sees all these as examples of the principle of complementarity, but he regards this principle as something of much broader scope and of deeper philosophical significance than as merely a principle limited in its application to physical systems. Thus he speaks of the impossibility of reconciling the demands of justice and mercy, and the presumptive impossibility of making a physical analysis of biological systems sufficiently searching to disclose the nature of life without destroying that life, as examples of the general principle of complementarity. It seems to me that it did not need quantum theory to disclose this general situation, but that we have always had situations where we have been forced to shift to another line of attack when we push our analysis to the logical limit. In other words, the method of "yes-but" we have always had with us. It seems to me that the generalized principle of complementarity is merely a glorified version of the principle of "yes-but". The method of "Yes-but" goes back at least to the time of Zeno, who, I will wager, was as capable as the next man of catching the tortoise which he intended to convert into stew for dinner, in spite of his paradoxes of motion. This sort of thing it seems to me is too ubiquitous and too vague to warrant our seeing here the operation of some grandiose "principle", nor do I

believe that it materially increases the presumptive truth of quantum theory to have discovered this sort of qualitative situation concealed in the consequences of its analysis. In fact, if it had not found this sort of thing it would be presumptive evidence against it. These strictures must not be taken as in any way reflecting on the validity of the numerical relationships demanded by quantum theory — these are an entirely different sort of thing.

Whatever view we take of complementarity as a grandiose principle of sweeping applicability, it seems obvious to me that here we have a factor militating against sharpness, for the line separating, for example, a legitimate space-time description from a deterministic description cannot be sharply drawn. Whenever we encounter such a lack of sharpness we may anticipate also a failure of the possibility of rigor.

All the situations which we have encountered thus far have a feature in common. In all of them we have encountered failures of our intellectual machinery to deal with experience as we obviously would like to have it deal — in particular, our intellectual machinery has proved itself incapable of exactly reproducing what we see happen. For instance, our verbalizing, or our mathematics, which is the same thing, has no built-in cut-off, corresponding to error or to the finiteness of human experience. In addition to this sort of failure of our mental machinery to exactly reproduce features of experience which are fairly obvious and which are often explicitly talked about, I think there is also failure for reasons not usually appreciated or said out loud, reasons corresponding to demands we *ought* to make of our mental machinery but which in fact we do not.

I think it will be admitted that an ideal mental machinery will not employ the operation of isolation for the reason that isolation does not occur in actuality. Quantum theory prohibits the isolation of the object of knowledge from the instrument of knowledge, and successfully analyzes the situations to which the Heisenberg principle applies in terms of the reaction between instrument and object which are ignored when they are isolated from each other. But any actual situation involves not only instrument of knowledge and object of knowledge, but also the knower. Quantum theory, however, consistently neglects the knower. Thus I find the following quotation in a recent lecture by Professor Bohr: 'In every field of experience we must retain a sharp distinction between the observer and the contents of the observations.' But in the world of things that happen this sort of distinction does not occur, and in making the distinction it seems to me that quantum theory practises a kind of isolation. In

physics the knower is always there, whether I am concerned with myself practising physics or whether I observe other people practising it. It may well be that quantum theory is justified for its particular purposes in taking the knower for granted, but we, in so far as we are committed to the problem of describing and understanding the total scene, may not neglect the knower. The problem of getting the knower into the picture has become acute now that most of us have become convinced that the knower is itself a physical system. Formerly, when people could think of mental activity as the functioning of a special mind stuff, *sui generis*, and with little in common with ordinary matter, it did not appear logically absurd to hope to give an account of the one kind of matter independently of the properties of the other. But now we are convinced that mental activity occurs in physical structures of stupendous complexity, made of the same atoms that the activity is seeking to comprehend. These complexities, if anything, increase the urgency of understanding the nature of the coupling between the structure of the brain and the external world. The presumption that there is some sort of essential limitation because of the nature of the structure and the coupling appears irresistible.

The concepts in terms of which we describe and understand the world about us do not occur in nature, but are man made products. Such things as length, or mass, or momentum, or energy occur only in conjunction with brains. The significance of these concepts cannot be isolated and associated only with the external world, but the significance is a joint significance involving external world and brain together. Now it seems to me that it is quite conceivable that different properties of the brain structure are involved in the concept of length, for example, than are involved in the concept of mass. It might be that the concept of mass is beyond the powers of certain simple types of brain whereas the concept of length might be easily within them. If such were the case, or if our present brain structure carries vestiges of limitations of this sort, our outlook might be materially altered.

To completely answer the questions brought up by considerations of this sort not only should we be able to hold ourselves to an awareness of the indissoluble tie-in of brain structure with the external world, but we should be able to describe specifically the nature of this tie-in for different concepts such as mass or length. We are at present hopelessly far from being able to do this, or even from knowing whether it is possible "in principle". There is, however, something which we can now do which has the effect of shifting the center of gravity away from the unknown

contribution of the brain, so that a little more "objectivity" can be imparted to our physical concepts. If one examines what he does when he determines such physical parameters of a physical system as its mass or its energy, it will be seen that the procedure involves the complicated interplay of operations of manipulation in the laboratory and operations of calculation. It is into these latter operations of calculation that the unknown and questionable influence of brain structure enters. Suppose now that we define the energy of a body, not as the number which is obtained by combining in a certain way other numbers which may correspond to velocity and mass, but as the number which is automatically given when a certain type of instrument, an "energy measurer" is coupled to the body. Such an "energy" is more something that we do and less something that we say and think than the conventionally defined energy. If we are clever we ought to be able to design instruments which would automatically record on a scale, when coupled to the body, any of the conventional physical parameters. When we have designed such instruments we should be able incidentally to discover some of the limitations in the measureability of energy, for instance, whereas it would be hopeless to expect to find such limitations as long as we have to treat the limitations as incidental to the structure of the brain.

I have made the beginning of an attempt to specify in detail how instruments might be constructed which would automatically register on a dial this or that physical parameter of an object when coupled to it. It is evident that the instruments will fall into hierarchies, the higher members of the hierarchy employing as component parts the complete instruments of lower levels. An instrument for automatically recording length is fairly easy to construct, whereas I found it to require great complication to construct an instrument for indicating mass, and even so there appear to be definite limitations on such features as speed of response. This is in spite of the fact that it is just as easy to *say* mass as to say length, and that in such an activity as dimensional analysis we think of mass and length as of equal simplicity. When we define mass and length instrumentally in this way, we see that, because of its greater complexity, it will not be so easy to apply the mass measuring instrument to small objects as the length measuring instrument, so that the concept of mass is subject to limitations in the direction of the very small to which the concept of length is not subject. This sort of limitation is entirely different from the sort of mutual limitation of measurements of velocity and position, for example, controlled by the Heisenberg principle in quantum

mechanics. It suggests itself that there may be other sorts of conceptual limitations in making contact with the world than those treated by conventional quantum mechanics.

The general problem back of all these later considerations is for the knower to know himself. It has been recognized as a fundamental philosophical problem since at least the time of Socrates, but it appears that we have not got very far toward a solution. Recent developments make it appear that the solution of this problem is more difficult than was perhaps at one time optimistically assumed. For we have here a self-reflexive situation, a system dealing with itself. Gödel's theorem shows that in the case of at least one special type of such a system there are drastic and formerly unsuspected limitations. It does not appear unreasonable to suppose by analogy that there are also formidable difficulties in the general case. I believe these difficulties appear the moment one attempts a specific attack on the problem — in fact it is difficult to even formulate what the problem is in self consistent language. It seems to me that nevertheless the problem is one of the very first importance. I think what I have said here makes it at least doubtful whether any possible solution can be rigorous in the canonical meaning of rigor — I believe that this will increase the difficulty of finding an acceptable solution rather than decrease it, as might perhaps at first seem natural. Until we have solved the problem, I do not believe that we can estimate what the limitations are on any possible rigor, nor even, for that matter, know what the true nature of rigor is.

LA FINITUDE EN MÉCANIQUE CLASSIQUE, SES AXIOMES ET LEURS IMPLICATIONS

ALEXANDRE FRODA

*Académie de la République Populaire Roumaine, Institut de Mathématiques, Bucarest,
Roumanie*

Le monde physique ne nous révèle jamais, du moins à notre échelle des grandeurs, l'existence actuelle de l'infini. En particulier, l'on ne rencontre en mécanique classique ni forces infinies, ni une infinité de renversements du sens de mouvement d'un mobile matériel en un laps fini de temps. C'est-ce qui nous a suggéré l'introduction de deux nouveaux axiomes en mécanique et l'étude de leurs implications. Nous les avons appelé axiomes de finitude (F).

À propos de la négation de l'infini (en mécanique classique) qui a inspiré ces axiomes, on peut citer une des profondes remarques de E. Mach sur l'évolution de la mécanique [8, 8; 34]: „Un des caractères de la connaissance instinctive”, écrivait-il, „c'est d'être surtout négative. Ce n'est pas prédire ce qui arrivera que nous pouvons faire, mais seulement dire les choses qui ne peuvent pas arriver, car celles-ci seules contrastent violemment avec la masse obscure des expériences, dans laquelle on ne discerne pas le fait isolé”.

On fera appel aux nouveaux axiomes afin d'éclaircir une question, qui s'est imposée à l'attention des physiciens, dès que Weierstrass prouva l'existence de fonctions continues sans dérivée. Or il est admis, en analyse, que les fonctions non dérivables ne sont nullement exceptionnelles dans la classe des fonctions continues. L'on admet, par contre, en mécanique classique que tout mouvement possède une vitesse et une accélération, à tout instant, ce qui implique l'existence des dérivées pour toutes les fonctions continues, qui définissent analytiquement les mouvements.

Soit

$$\vec{r} = \vec{r}(t) \quad (1)$$

une équation vectorielle définissant la cinématique du mouvement μ d'un point matériel M de masse m , dans un laps de temps $\delta = [t_0, t_1]$. Il y est supposé, selon Newton, que t mesure physiquement, à partir d'un instant

initial $t = t_0$, le temps „absolu” et que le vecteur de position \vec{r} , situe le point M par rapport à un système fixe d'axes cartésiennes constituant des repères de l'espace „absolu”.

L'on pose, en mécanique classique, pour la vitesse et l'accélération du mobile à l'instant t ,

$$\vec{v}(t) = \frac{d}{dt} \vec{r}(t), \quad \vec{A}(t) = \frac{d}{dt} \vec{v}(t), \quad (2)$$

ce qui les définit aussi comme fonctions vectorielles de t .

On y admet la continuité de $\vec{r}(t)$, qui résulte de notre intuition du temps et du mouvement. Cette assertion ne sera pas mise en discussion, à cette occasion.

La définition (2) de $\vec{v}(t)$ a un sens en mécanique classique par ce que la fonction vectorielle $\vec{r}(t)$ y est supposée dérivable, propriété attribuée à tout mouvement. Afin de justifier la définition (2) de $\vec{A}(t)$, il y est admis, de plus, que $\vec{v}(t)$ est non seulement continu, mais aussi dérivable, par rapport à t , quel que soit $t \in \delta$.

Ainsi tout mouvement μ présenterait en mécanique classique des caractères, qui ne sont démontrables „ni mathématiquement, ni empiriquement”, comme l'affirmait G. Hamel [5a; 5b, p. 64; 5c, p. 2] en axiomatisant la mécanique rationnelle. En faisant remarquer, que „aucune expérience ne serait assez fine pour descendre jusqu'au différentielles”, Hamel attribuait l'existence de la vitesse et de l'accélération d'un mouvement μ , à tout instant t de δ , à un principe physique, selon lequel: „Toutes les grandeurs observables sont continues et continument différentiables”. Un principe pareil fut affirmé par L. Zorretti [10, pp. 16, 17, 40], dans son étude des principes de la mécanique classique.

Désignons par Ra un axiome affirmant l'existence d'une accélération $\vec{A}(t)$ à tout instant t d'un mouvement μ , du domaine C de valabilité de la mécanique classique (de Newton), ce qui implique aussi l'existence d'une vitesse $\vec{v}(t)$ continue.

Il nous faudra distinguer entre les mouvements à définition purement cinématique (1) et les mouvements μ_c réalisables en C , ce qui exige des définitions explicites. En faisant abstraction des éventuelles résistances passives (frottement viscosité, etc.) d'un mouvement réel μ de C , l'on

fait correspondre à μ un mouvement μ_c „conservatif”¹, qui est soit égal à μ , soit — lorsque μ n'est pas conservatif — égal à la limite (cinématique) de la suite des mouvements non-conservatifs obtenus en faisant tendre successivement les résistances passives de μ vers zéro. Cela est considéré possible, sinon expérimentalement, du moins théoriquement.

Soit (R) un système classique d'axiomes de la mécanique rationnelle. Tout mouvement réalisable en C y satisfait, mais la réciproque pourrait ne pas être vraie. En effet, un mouvement quelconque d'un point matériel M de masse m étant donné par l'équation (1), il semble douteux qu'un tel mouvement soit réalisable, quelle que fût sa définition cinématique. C'est un fait que signalait déjà H. Herz [7, p. 12], dans son étude des principes de la mécanique. Nous aurons à revenir tout à l'heure sur ce point tout aussi important, que délicat.

Le domaine C des mouvements, considérés en mécanique rationnelle du temps de Newton, fut ultérieurement réduit par suite de la critique des principes qu'il avait posé à la base de sa „philosophie naturelle”, critique stimulée par les progrès ultérieurs de la physique. Newton avait admis à la fois les principes suivants: 1) l'existence d'un temps et d'un espace „absolus”, ainsi que la constance „absolue” de la masse en mouvement et 2) Le maintien des propriétés de la matière jusque dans ses parties ultimes, „indivisibles”. De ces principes, le premier est aujourd'hui contesté par la mécanique de la relativité générale, le second par la mécanique quantique (ondulatoire). En conséquence, le domaine C de la mécanique rationnelle est limité aujourd'hui par l'existence de ces dernières mécaniques. C'est pourquoi nos axiomes de finitude s'appliquent seulement à la mécanique rationnelle, sans préjudice de leur éventuelle extension aux mécaniques nouvelles. Or, de la mécanique du point on est conduit à la mécanique des systèmes en vertu d'axiomes que nous n'allons pas examiner.

Signalons toutefois que la notion de point matériel pose elle-même des questions. Depuis Euler et Lagrange l'on admet souvent en mécanique classique, l'existence de points matériels M aux dimensions nulles, mais de masse m non-nulle. L. Zorretti [10] les appelle „fictifs”, puisque l'on y néglige les propriétés rotationnelles d'un corps très petit. Mais il y a plus, l'introduction de tels points peut conduire à des contra-

¹ Un mouvement sera dit *conservatif*, par définition, s'il ne comporte pas de dégradation d'énergie (due à des résistances passives). Pour le sens différent, classique, attribué à l'expression „système conservatif”, voir par exemple Appell *P., Traité de mécanique rationnelle*, t. II, Ed. 4, Paris (1923), p. 65 et suivantes.

dictions. Considérons par exemple, abstraction faite des résistances passives, le mouvement de M le long d'une courbe Γ matérielle, plane, verticale, d'équation $y = x^{4/3}$, où Oy est la verticale. Si M doit être pressé contre Γ , de son côté concave, il faut que M soit un point géométrique, puisque le rayon de courbure ρ égale zéro au sommet de Γ . Or si la vitesse $\rightarrow v$ n'y était pas nulle, la force de liaison y serait infinie, ce qui n'est pas physiquement réalisable.

L'on peut remarquer d'ailleurs, que l'existence d'un point matériel sans dimensions est tout aussi critiquable, que l'admission d'existence d'un instant réel t de temps, à durée nulle, qui peut, de même, conduire à des contradictions. Ces existences physiquement inconcevables sont parfois impliquées par l'application de la méthode infinitésimale en physique, qui devrait — semble-t-il — être l'objet d'une analyse axiomatique, assez difficile à faire.

De telles objections apparaissent aussi dans les mécaniques nouvelles. Signalons ainsi, en passant, un passage significatif où Heisenberg en s'occupant de son principe d'incertitude exprimait, déjà en 1930, des doutes de principe sur la légitimité d'attribuer un sens physique au passage à la limite d'un volume et d'une durée élémentaires, lorsqu'il s'agit d'évaluer l'amplitude d'un champ électrique et d'un champ magnétique en mécanique quantique [6, p. 37].

Revenons à notre problème, qui est celui de débarrasser les axiomes de la mécanique classique de l'hypothèse Ra , définie ci-dessus. L'on y parviendra en considérant d'abord dans les conditions les plus générales de l'analyse les grandeurs vectorielles, qui interviennent en mécanique et en recherchant ensuite les circonstances, qui imposent l'existence des dérivées, lorsque les mouvements sont réalisables en C . C'est ce que nous avons entrepris dans un travail antérieur, en roumain [3, p. 3-4].

Le développement du programme indiqué exige des notions de cinématique générale, que l'on définit en étendant aux fonctions vectorielles $W(t)$ d'une variable réelle t les propriétés classiques des fonctions (numériques) réelles de t , concernant la continuité, les bornes, les limites, la dérivation et l'intégration. Il nous suffit d'en mentionner l'analogie, que reflète la terminologie respective.

Voici enfin quelques définitions de cinématique générale, qui nous serviront à formuler les axiomes de la mécanique rationnelle. Considérons, de nouveau, un mouvement μ à définition cinématique (1). Par définition, la \rightarrow vitesse $v(t)$ existe à l'instant t , si pour $\Delta t \rightarrow 0$, cela a un sens d'écrire

$$\vec{v}(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} [\vec{r}(t + \Delta t) - \vec{r}(t)] \quad (3)$$

et cette vitesse sera dite *complète*. De même, il existe à l'instant t une vitesse *prospective* $\vec{v}_+(t)$ ou *rétrospective* $\vec{v}_-(t)$, lorsque la limite en (3) a un sens pour $\Delta t > 0$ ou pour $\Delta t < 0$, respectivement. Considérons, en particulier, le cas d'un mouvement μ , tel que $\vec{v}(t)$ existe et soit continue à chaque instant t de $\delta[t_0, t_1]$. Par définition, l'accélération $\vec{A}(t)$ existe alors à l'instant t , si pour $\Delta t \rightarrow 0$, cela a un sens d'écrire

$$\vec{A}(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} [\vec{v}(t + \Delta t) - \vec{v}(t)] \quad (4)$$

et cette accélération sera dite *complète*. De même, il existe à l'instant t une accélération *prospective* $\vec{a}(t)$ ou *rétrospective* $\vec{\alpha}(t)$, lorsque la limite en (4) a un sens pour $\Delta t > 0$ ou pour $\Delta t < 0$, respectivement.

Nous verrons, que dans le domaine C de la mécanique rationnelle la vitesse complète $\vec{v}(t)$ existe à tout instant d'un mouvement réalisable en C et qu'il y a donc alors

$$\vec{v}(t) = \vec{v}_+(t) = \vec{v}_-(t).$$

C'est pourquoi il est inutile de définir des „accélérations”, en partant des vitesses prospective $\vec{v}_+(t)$ et rétrospective $\vec{v}_-(t)$.

Aux notions cinématiques précédentes ajoutons encore les définitions suivantes, afin d'abrégier le langage:

1) On dira qu'un mouvement μ , à définition cinématique (1) est *régulier*, en un laps de temps $\delta = [t_0, t_1]$, s'il possède à chaque instant t de δ une accélération complète $\vec{A}(t)$ continue.

2) En considérant les mouvements μ d'équation $\vec{r} = \vec{r}(t)$ et μ_n d'équations $\vec{r} = \vec{r}_n(t)$, $n = 1, 2, \dots$, l'on dira que μ est la *limite cinématique* des μ_n , pour $n \rightarrow \infty$, lorsque $\vec{r}(t) = \lim_{n \rightarrow \infty} \vec{r}_n(t)$.

3) On dira qu'un vecteur $\vec{W}(t)$ change d'orientation dans l'espace une infinité de fois en une suite indéfinie σ (croissante, resp. décroissante) d'instants successifs t_i , lorsqu'il existe un axe de l'espace, tel que les pro-

jections des $\vec{W}(t_i)$, $\vec{W}(t_{i+1})$ sur cet axe aient des sens opposés, pour une infinité d'indices parmi les $i = 1, 2, \dots$

4) Lorsque dans un intervalle $\delta = [t_0, t_1]$ des mouvements μ_1, μ_2 sont donnés par $\vec{r} = \vec{r}_1(t)$, $\vec{r} = \vec{r}_2(t)$, l'on dira que le mouvement μ donné par $\vec{r} = \vec{r}(t)$ est leur *mouvement résultant*, s'il y a $\vec{r}(t) = \vec{r}_1(t) + \vec{r}_2(t)$.

5) Appelons *polynomial* un mouvement $\vec{r} = \vec{r}(t)$, où $\vec{r}(t)$ est un polynôme en t , dont les coefficients sont des vecteurs constants de l'espace. Un tel mouvement est régulier.

Ajoutons la remarque suivante: Les fonctions vectorielles continues de t étant des limites (uniformes) de polynômes, tout mouvement μ , à définition cinématique (1), d'un point M , peut s'exprimer (et de bien de manières différentes) comme limite cinématique d'une suite de mouvements polynomiaux μ_n du même point.

Revenons maintenant au système classique (R) des axiomes de la mécanique. Il est admis, en mécanique rationnelle, que la force \vec{F} qui produit le mouvement μ d'un point matériel M doit exister à tout instant t de δ (même si $\vec{F} = 0$) et que μ est soumis, à cet instant, à la loi de Newton (axiome Rn), qui s'écrit

$$\vec{F} = m \cdot \vec{I}, \quad (5)$$

où m est la masse *constante* de M et \vec{I} son accélération à l'instant t . E. Mach et P. Painlevé n'ont vu en Rn , qu'une définition de la force \vec{F} qui produit le mouvement, mais G. Hamel y reconnut une relation effective, car il existe des classes (Φ) de phénomènes physiques, telles qu'il y ait, pour chacune d'elles, lorsque m désigne la masse constante du point matériel M une loi générale

$$\vec{F} = m \cdot \vec{\Phi}(\vec{r}, \vec{v}, t),$$

où $\vec{\Phi}$ est une fonction vectorielle des variables \vec{r}, \vec{v}, t attachée à la classe (Φ) — et y constituant bien souvent le vecteur d'un champ. S'il est question d'un mouvement μ *déterminé*, réperé par (1) et tel que $\vec{v}(t)$ existe en δ , les formules (3), (4) montrent que, pendant le mouvement, il y a $\vec{F} = \vec{F}(t)$,

de sorte que la force qui produit μ est, en ce cas, égale à une fonction vectorielle de t et seulement de t .

L'existence d'une telle force, produisant un mouvement déterminé, réalisable en C est assurée par un axiome Rd . On peut prouver que l'accélération \vec{F} , dont l'existence est admise en (5) n'y intervient que sous la forme d'accélération prospective $\vec{a}(t)$:

1) Voici un premier argument: S'il y avait $\vec{F} = \vec{A}$ pour un mouvement μ réalisable en C et à tout instant t de δ , les discontinuités de $\vec{F}(t)$ devaient avoir, en vertu de (5), les caractères d'une dérivée vectorielle $\vec{A}(t)$ et ne pourraient être donc de première espèce (c.-à-d. présenter un saut). Or cela est contredit par des indications claires de l'expérience physique, comme le montrent les exemples suivants:

a. Considérons la force discontinue, qui met en mouvement le poids équilibré de la machine d'Atwood, à l'instant d'arrêt de la masse additionnelle. Cette force exécute un saut.

b. Considérons la force discontinue, qui produit le mouvement d'un point soumis à l'attraction newtonienne d'une surface sphérique fermée, à l'instant où il la traverserait. Cette force exécute un saut.

2) Voici un second argument: Lorsque l'action d'une force cesse de s'exercer ($\vec{F} = 0$) sur un point matériel M , il continue à se déplacer d'un mouvement rectiligne et uniforme en vertu de sa vitesse acquise. Or si l'on avait, en (5), l'égalité $\vec{F} = \vec{A}(t)$, à chaque instant t de δ , la force \vec{F} serait prédéterminée à l'instant t_0 par le mouvement antérieur à t_0 , puisque l'accélération $\vec{A}(t_0) = \vec{a}(t_0)$ est donnée par les valeurs de $\vec{r}(t)$, pour $t \leq t_0$. Cela est non seulement paradoxal, mais devient évidemment absurde, lorsque $\vec{F}(t)$ est discontinue pour $t = t_0$ et de plus, contredit la conception d'une force, cause de modification du mouvement d'inertie, dont l'existence est assurée par un axiome Ri .

Citons aussi deux axiomes de (R) , se complétant l'un l'autre: L'axiome Rc affirme que le mouvement résultant (cinématique) de mouvements réalisables en C est aussi réalisable en C et l'axiome Rf affirme que sous l'action simultanée de plusieurs forces, c'est la force égale à leur résultante vectorielle, qui les remplace.

Une question délicate, déjà signalée en passant, est la suivante: Le système (R) des axiomes classiques, y inclus Ra , est-il aussi une condition

suffisante pour que tout mouvement, défini cinématiquement fût aussi réalisable en C ? Cela n'est pas du tout vraisemblable et il est même douteux, qu'il puisse exister un système d'axiomes représentant des conditions nécessaires et suffisantes afin que tout mouvement μ les satisfaisant fût réalisable en C .

Nous pouvons énoncer pourtant des conditions suffisantes pour que certains mouvements soient réalisables en C . Les voici, sous forme de propositions que l'on peut démontrer *sans faire appel aux axiomes de la mécanique rationnelle*; mais par une méthode constructive:

I. Tout mouvement polynomial d'un point matériel M est réalisable en C .

II. Tout mouvement μ est réalisable en C en même temps que ses projections μ_x, μ_y, μ_z sur les axes.

III. Lorsqu'un mouvement μ d'un point matériel M , défini cinématiquement en un laps de temps $\delta = [t_0, t_1]$ possède une accélération $\rightarrow a(t)$ prospective continue et qui ne change pas une infinité de fois son orientation dans l'espace, le mouvement μ est réalisable en C .

Les démonstrations des propositions I, II et III consistent à mettre en évidence la possibilité de construire le mouvement μ respectif à l'aide de mécanismes convenables, quand on fait abstraction des résistances passives. Ces constructions jouent le rôle de modèles existentiels.

Nous pouvons énoncer aussi des conditions nécessaires à ce qu'un mouvement soit réalisable en C , en complétant d'une part le système (R) avec les axiomes de finitude (F) , tout en abandonnant d'autre part l'axiome Ra , qui affirmait *a priori* l'existence de l'accélération à tout instant d'un mouvement réalisable en C .

Voici enfin l'énoncé de nos axiomes de finitude, valables en mécanique rationnelle.

(F) . Lorsqu'un mouvement μ est réalisable en C , dans un laps de temps δ , fini, il satisfait aux conditions:

$F1$. Parmi les suites de mouvements réguliers, réalisables, ayant pour limite cinématique le mouvement μ , il existe une suite de mouvements $\mu_{n=1,2,\dots}$, tels que les forces $\rightarrow F_n$ qui les produisent soient bornées dans leur ensemble.

$F2$. La force $\rightarrow F(t)$, qui produit le mouvement μ en δ ne peut changer d'orientation une infinité de fois, en aucune suite indéfinie σ (croissante, resp. décroissante) d'instants successifs.

On doit remarquer que l'adjonction des axiomes de finitude (F) à un système classique (R) d'axiomes du mouvement en C doit être effectuée,

en même temps que l'abandon de l'axiome *Ra*. Mais l'on ne peut renoncer à l'hypothèse d'existence de l'accélération, qu'en modifiant à la fois l'expression d'autres axiomes de (*R*), afin de ne plus admettre explicitement

l'existence des $\vec{v}(t)$ et $\vec{A}(t)$, pour tout $t \in \delta$. Le nouveau système (\bar{R}) d'axiomes, ainsi obtenu, remplacera (*R*) et nous allons en exposer les principales implications, où le rôle des axiomes (*F*) est essentiel.

Afin de les obtenir, on s'appuiera sur des propriétés générales des fonctions vectorielles de *t*, ainsi que sur les propositions I) II) III) ci-dessus, qui expriment des conditions suffisantes, afin que certains mouvements soient réalisables en *C*. L'on obtient les résultats suivants, qui expriment des *propriétés appartenant à tout mouvement μ réalisable en C*:

1°. Il existe à chaque instant *t* de δ une vitesse complète $\vec{v}(t)$ continue et des accélérations prospective $\vec{a}(t)$ et rétrospective $\vec{\alpha}(t)$.

2°. L'accélération prospective $\vec{a}(t)$ est prospectivement continue et ne possède qu'un nombre fini d'instants de discontinuité en un laps δ fini.

3°. L'accélération complète $\vec{A}(t)$ existe et est continue à chaque instant *t* de δ , sauf en un nombre fini (ou nul) d'instants t_k de δ , où $\vec{a}(t)$, $\vec{\alpha}(t)$ sont discontinues et $\vec{a}(t_k) \neq \vec{\alpha}(t_k)$.

4°. *Tout μ est, en δ , soit un mouvement régulier, soit une succession finie de mouvements réguliers.*

La démonstration des propriétés précédentes utilise l'appareil mathématique de la théorie des fonctions vectorielles de variables réelles [1]. A part quelques propositions connues, où qui étendent directement aux fonctions vectorielles des propriétés classiques des fonctions numériques de variable réelle, nous avons fait appel bien souvent à une proposition inspirée par ces recherches même et que voici:

„Lorsque parmi les vecteurs dérivés prospectifs (resp. rétrospectifs) d'une fonction vectorielle $\vec{V}(t)$ pour $t = t_0$, fonction possédant des vecteurs dérivés, bornés dans leur ensemble en $\delta = [t_1, t_2]$, $t_0 \in \delta$, il y a deux vecteurs dérivés \vec{D}_1, \vec{D}_2 , faisant entre eux un angle non-nul, il existe un vecteur variable $\vec{W}(\tau_p)$ égal à la dérivée vectorielle unique

$$\vec{W}(t) = \frac{d}{dt} \vec{V}(t)$$

pour $t = \tau_p$ et qui change son orientation dans l'espace une infinité de fois, dans une suite de valeurs τ_p , $p = 1, 2, \dots$, tendant vers t_0 en décroissant (resp. en croissant)".

J'ajoute, que — par définition — un „vecteur dérivé prospectif" du vecteur variable $\vec{V}(t)$, pour $t = t_0$, correspond, par analogie, à l'un des coefficients différentiels d'une fonction réelle $f(t)$ à droite, tandis que la „dérivée vectorielle" correspond à la dérivée unique, pour $t = t_0$.

Les résultats précédents de 1° à 4° font directement appel aux axiomes, notés précédemment par Rd , Ri , Rn , Rc , Rf , $F1$, $F2$ et n'utilisent pas l'axiome Ra . Les autres axiomes Rr , sur l'égalité de l'action et de la réaction et Ru , qui assure l'unicité d'un mouvement réalisable en C , pour des conditions initiales données sous l'action de forces données, n'y interviennent pas, du moins explicitement. Or on a vu que certains axiomes de (\vec{R}) doivent être exprimés sous une forme modifiée, avant de former avec les axiomes (F) le nouveau système (\vec{R}) . Voici des exemples, qui font apparaître les modifications en question :

1°. Axiome Ri (loi d'inertie) : „Lorsqu'à un instant initial t_0 du laps δ , un point matériel M possède une vitesse rétrospective $\vec{v}_-(t_0)$ et qu'aucune force ne s'exerce sur lui pendant δ , le point M décrit en δ un mouvement rectiligne de vitesse $\vec{v}(t)$ constamment égale à $\vec{v}_-(t_0)$ ".

En usant de la vitesse *rétrospective* initiale, l'on évite de réintroduire (même sous une forme affaiblie) l'hypothèse d'existence de la vitesse et il suffit en effet, d'admettre physiquement, qu'on dispose du mouvement de M dans un laps de temps, aussi petit qu'on veut, antérieur à t_0 .

2°. Axiome $\vec{R}n$ (loi de Newton) : „Si dans un mouvement μ réalisable en C d'un point matériel M , possédant à chaque instant $t \in \delta$ une vitesse $\vec{v}(t)$ continue, il existe à un certain instant $t_1 \in \delta$ une accélération prospective \vec{a} et si la force, qui s'exerce sur M est \vec{F} , il y a

$$\vec{F} = m \cdot \vec{a},$$

où m est une constante, dépendant de M et indépendante du mouvement μ -à- d . de t ."

Il est clair que l'existence de l'accélération prospective $\vec{a} = \vec{a}(t_1)$ n'est admise, dans cet énoncé, que pour l'instant $t = t_1$.

Les axiomes de finitude (F) ne prétendent pas à être acceptés, comme l'axiome Ra , sans confronter l'expérience. L'on peut concevoir des ex-

périences dont les résultats prévisibles constituent une vérification de ces axiomes. Voici le schéma d'une de ces expériences, ayant lieu sans résistances passives et utilisant des solides parfaitement élastiques :

Soit un petit pendule simple vertical P , dont les oscillations sont limitées de chaque côté par des obstacles plans, verticaux, symétriques par rapport au plan vertical V contenant l'axe de suspension. Ces obstacles sont reliés au sol, de manière à posséder des mouvements uniformes, autonomes, indépendants des chocs du petit pendule et tels qu'ils arrivent simultanément à l'instant t_1 en V . Selon les lois classiques du choc le pendule devrait effectuer une infinité de demi-oscillations, d'amplitudes décroissantes, en un laps $\delta = [t_0, t_1]$, ce qui contredirait $F2$. Or, en réalité, le nombre d'oscillations, ne peut être que fini, ce qui se vérifie aisément, si l'on tient compte de la durée des chocs, calculable selon la loi de Herz. C'est pourquoi l'axiome $F2$ sera vérifié par cette expérience. Négligier la durée des chocs engendre des paradoxes, comme celui remarqué par D. Gale [4], qui pensait avoir signalé un cas d'indétermination en mécanique classique.

On peut aussi établir par le raisonnement l'indépendance des axiomes (F) par rapport au système classique $\{(R) - (Ra)\}$ d'axiomes.

La question, que pose l'extension éventuelle des axiomes de finitude, valables en C , aux phénomènes étudiés par les mécaniques nouvelles est un problème ouvert.

On peut rapporter à ce problème quelques faits bien connus, élémentaires, qu'on peut relier aux axiomes (F) et à leurs implications en C :

1) Dans la mécanique de la relativité générale, où la masse est fonction de la vitesse du mouvement, il y a la vitesse c de la lumière, qui pose une borne finie à la vitesse de tout mouvement, ce qui doit affecter l'expression de l'axiome $F1$.

2) En mécanique quantique l'on se rappelle que l'existence d'une vitesse a été mise en doute, dès les premières études du mouvement brownien. Ainsi, en expérimentant, J. Perrin signalait une analogie d'aspect du mouvement brownien aux fonctions sans dérivées [9, p. 164], ce qui confirmait les vues de Einstein, lequel, dans ses études théoriques, avait démontré auparavant que la vitesse moyenne en Δt du mouvement d'une particule ne tend vers aucune limite, lorsque la durée Δt tend vers zéro. Il concluait en faisant remarquer que pour l'observateur de ce mouvement la vitesse moyenne lui apparaîtrait comme vitesse instantanée, mais qu'en fait elle ne représente aucune propriété objective du mouvement soumis à

l'investigation, *du moins si la théorie correspond aux faits*, ajoutait-il [2].
Par ces paroles d'extrême prudence, je conclus aussi mon exposé.

Bibliographie

- [1] BOURBAKI, N., *Fonctions d'une variable réelle*. Livre IV, Chap. I, II, III, Paris 1949.
- [2] EINSTEIN, A., *Zur Theorie der Brownschen Bewegung*. Annalen der Physik, Serie 4, Vol. 19 (1906), pp. 371–381.
- [3] FRODA, A., *Sur les fondements de la mécanique des mouvements réalisables du point matériel* (en roumain). Studii și Cercetări Matematice, t. III, București 1952.
- [4] GALE, D., *An indeterminate problem in classical mechanics*. Amer. Math. Monthly, vol. 59 (1952), pp. 291–295.
- [5] HAMEL, G., a) *Ueber die Grundlagen der Mechanik*. Math. Annalen, Bd. 66 (1908), pp. 350–397.
b) *Elementare Mechanik*. Leipzig, 1912.
c) *Die Axiome der Mechanik*. Handbuch der Physik, Bd. 5, Berlin 1927, pp. 1–42.
- [6] HEISENBERG, W., *Die Physikalischen Prinzipien der Quanten-theorie*. Leipzig 1941.
- [7] HERZ, H., *Die Prinzipien der Mechanik in neuem zusammenhange dargestellt, Gesammelte Werke*. Bd. III, Leipzig 1910.
- [8] MACH, E., *La Mécanique, exposé historique et critique de son développement* (trad. Em. Bertrand). Paris 1904.
- [9] PERRIN, J., *L'Atome*. Paris 1912.
- [10] ZORETTI, L., *Les principes de la mécanique classique*. Mémorial des Sciences Math., Paris 1928.

THE FOUNDATIONS OF RIGID BODY MECHANICS AND THE DERIVATION OF ITS LAWS FROM THOSE OF PARTICLE MECHANICS

ERNEST W. ADAMS

University of California, Berkeley, California, U.S.A.

1. **Introduction.** This paper has three purposes: (1) to give a system of axioms for classical rigid body mechanics (henceforth abbreviated '*RBM*'); (2) to show how these axioms can be derived from those of particle mechanics (abbr. '*PM*'); and (3), using the foregoing derivation as an example, to give a general characterization of the notion of '*reduction*' of theories in the natural sciences. The axioms to be given are due jointly to Herman Rubin and the author. They comprise what may be thought of as the theory of rigid motions under finite applied forces with moments of inertia given. That part of *RBM* which deals with the calculation of moments of inertia from known mass distributions is omitted, since the laws of motion can be stated directly in terms of total masses and moments of inertia. Similarly, the theory of impacts, which cannot be represented in terms of finite forces, is excluded. In the axioms, the laws of rigid motion are presented *de novo*, and are not, as is usually the case, presented as deductive consequences of the laws of *PM*. It is our contention that, in spite of superficial differences from the more well-known examples, the derivation of the laws of *RBM* from those of *PM* can be viewed as an example of reduction. In section 3 we shall analyse the logical relation which must hold between two theories in order that one should be reduced to the other, and then in the final section we shall indicate how *RBM* may be reduced to *PM* in accordance with the theory of reduction previously given.

Because of limitations of space, our discussion both of the theories of *RBM* and *PM* and of the general concept of reduction and its specific application to *RBM* and *PM* will be limited. A complete formal development of these topics is given in the author's Ph. D. dissertation, *The Foundations of Rigid Body Mechanics* [1].

2. **Axioms of Rigid Body Mechanics.** Our axioms are based on seven primitive notions, five of which are closely analagous to the primitive

notions of McKinsey, Sugar, and Suppes' axiomatization of classical *PM* [5]. These seven are denoted ' K ', ' T ', ' g ', ' R ', ' H ', ' μ ', and ' 0 ', and their intended interpretations are as follows:

K is a set of *rigid bodies*.

T is an interval of real numbers representing *clock readings* during an interval of time.

g is a function from K to the positive real numbers, such that for every rigid body k in K , $g(k)$ is the *mass* of k as measured in some fixed units.

R is a function from $K \times T$ to E_r (the set of ordered r -tuples of real numbers) such that for each k in K and t in T , $R(k, t)$ is the r -vector representing the *position of the center of mass* of k at the instant when the clock reads t , as measured relative to a system of cartesian coordinate axes. r -vectors are here construed to be ordered r -tuples of real numbers, and, of course, in the ordinary application, $r = 3$.

H is a function from $K \times T \times N$ (N being the set of positive integers) to $E_r \times E_r$, such that $H(k, t, n)$ represents the n 'th *applied force* acting on body k at the time t in the following way: $H(k, t, n)$ is the r -vector representing the magnitude and direction of the n 'th applied force, and $H^2(k, t, n)$ is the r -vector representing the position of the point of application of this force relative to a specially selected system of coordinate axes which are parallel to the original reference frame, but which have their origin at the center of mass of k . We shall call the original axes the '*axes of the space*,' and the new axes the '*non-rotating axes of k* '.

μ is a function from K to the set of r by r matrices with real components, such that for each k in K , $\mu(k)$ is a matrix representing the *moment of inertia* tensor of k relative to still another set of coordinate axes, which we shall call the '*rotating axes of k* .' The rotating axes of k are a system of cartesian coordinate axes which have, like the non-rotating axes of k , their origin at the center of mass of k , but which rotate with k so that they always maintain a fixed relation to the parts of k . If k is composed of a finite number of mass points with masses m_1, \dots, m_i , and positions L_1, \dots, L_i relative to the rotating axes of k , then the matrix $\mu(k)$ is the sum of the products:

$$\mu(k) = \sum_{j=1}^i m_j(L_j)^*(L_j).^1$$

¹ The *transpose* L^* of an r -vector L is a 'column vector' with r rows, and the dyadic product of a column vector L^* and a row vector, say M (both r -vectors) is

The matrix $\mu(k)$ defined as above is symmetric and positive semi-definite since all of the m_j 's are positive. It is to be particularly noted that moment of inertia, as characterized here, is independent of time, because of the fact that it is defined relative to the rotating axes of k , which always remain fixed within k . To transform to the time-dependent moment of inertia function used in many formulations of the laws of *RBM* (e.g., Milne [7], p. 267 or Joos [3], p. 137 or McConnell [4], p. 233), it is necessary to introduce our last primitive notion, a function representing the orientation in space of the rotating co-ordinate axes of k .

0 is a function from $K \times T$ to the set of r by r orthogonal matrices, such that for each k in K and t in T , $\theta(k, t)$ represents the 'orientation' of the set of rotating coordinates of k at time t relative to the axes of the space. $\theta(k, t)$ gives the orientation of the moving axes in the sense that for each $j = 1, \dots, r$, $\theta(k, t)_j$ — the j 'th row of the matrix $\theta(k, t)$ — is the unit vector in the direction of the j 'th axis of the moving axes of k at time t . Or, $\theta(k, t)_{i,j}$ is the cosine of the angle between the i 'th rotating axis of k at time t and the j 'th axis of the space.

The equation which relates the time dependent moment of inertia function $\mu(k, t)$ and the time-independent moment of inertia function μ is simply:

$$\mu(k, t) = \theta^*(k, t)\mu(k)\theta(k, t).$$

The axioms for *RBM* can now be stated in terms of the seven primitive concepts just discussed. The style in which these axioms are formulated is very similar to that of the axioms for classical particle mechanics due to McKinsey, Sugar, and Suppes [5], and the axioms for relativistic particle mechanics due to Rubin and Suppes [10] (see also, McKinsey and Suppes [6]). That is, the axioms are conditions which are parts of the definition of the set-theoretical predicate system of *r-dimensional rigid body mechanics*. Our axioms rely directly on the concept of a system of *r-dimensional particle mechanics*, which is defined as follows:

DEFINITION 1. An ordered quintuple $\langle P, T, m, S, F \rangle$ is a SYSTEM OF CLASSICAL *r-DIMENSIONAL PARTICLE MECHANICS* if and only if it satisfies axioms P1–P6.

P1. P is a non-empty finite set.

P2. T is an interval of real numbers.

an r by r matrix $(L)^*(M)$ such that the element of its i 'th row and j 'th column is $[(L)^*(M)]_{i,j} = L_i M_j$.

- P3. S is an r -vector valued function with domain $P \times T$ such that for all p in P and t in T , $d^2/dt^2(S(p, t))$ exists.
- P4. m is a positive real-valued function with domain P .
- P5. F is an r -vector valued function with domain $P \times T \times N$, where N is the set of positive integers, and for all p in P and t in T the series $\sum_{i=1}^{\infty} F(p, t, i)$ is absolutely convergent.
- P6. For all p in P and t in T ,

$$m(p) \frac{d^2}{dt^2} (S(p, t)) = \sum_{i=1}^{\infty} F(p, t, i).$$

In the above axioms, P is to be thought of intuitively as a set of particles, T — again — is an interval of clock readings, $m(p)$ is the mass of particle p , $S(p, t)$ is the r -vector representing the position of p at time t , relative to a system of cartesian coordinate axes, and $F(p, t, i)$ is an r -vector representing the magnitude and direction of the i 'th force applied to p at time t (in the case of particle mechanics it is not necessary to take into account the point of application of a force since this affects only the rotation, and not the translation of a particle). The only axiom embodying what is normally thought of as a 'physical law' is P6, expressing a version of Newton's Second Law. The first five axioms serve only to define the set-theoretical character of the primitive notions, and state certain continuity and differentiability conditions.

The axioms for *RBM* are stated in Definition 2, below. It will be seen that only two of them contain ordinary physical laws, and the remainder, like axioms P1–P5 in Definition 1, stipulate the set-theoretical type of the primitives. Axiom R1, stating that the first five elements of a system of *RBM* are themselves a system of *PM*, contains Newton's Second Law, since the axioms for *PM* include this law; and axiom R5 is a version of well-known tensor equations relating moment of inertia, angular acceleration (these two being combined to give the rate of change of angular momentum), and resultant torque, or moment force.

DEFINITION 2. *An ordered septuple $\langle K, T, g, R, H, \mu, 0 \rangle$ is a SYSTEM OF r -DIMENSIONAL RIGID BODY MECHANICS if and only if satisfies axioms R1–R5.*

- R1. H is a function with domain $K \times T \times N$ taking as values ordered pairs of r -vectors, and if H^1 and H^2 are r -vector valued functions with

domain $K \times T \times N$ such that for all k in K , t in T and i in N ,

$$H(k, t, i) = \langle H^1(k, t, i), H^2(k, t, i) \rangle,$$

then $\langle K, T, g, R, H^1 \rangle$ is a system of classical r -dimensional particle mechanics.

- R2. θ is a function with domain $K \times T$ taking as values r by r orthogonal matrices, such that for all k in K and t in T , $d^2/dt^2(\theta(k, t))$ exists.
- R3. μ is a function with domain $K \times T$ taking as values r by r symmetric positive semi-definite matrices of rank r or $r - 1$.
- R4. For all k in K and t in T , the series

$$\sum_{i=1}^{\infty} H^2(k, t, i) \times H^1(k, t, i) \text{ is absolutely convergent. }^2$$

- R5. For all k in K and t in T ,

$$\theta(k, t) \times \left[\mu(k) \frac{d^2}{dt^2} (\theta(k, t)) \right] = \sum_{i=1}^{\infty} H^2(k, t, i) \times H^1(k, t, i).$$

The axioms of Definition 2 are all of fairly simple significance. R1 states essentially that the system which is formed by taking only the masses, positions of the centers of mass of the rigid bodies, and the magnitudes and directions of the applied forces, constitutes a system of particle mechanics; i.e. it obeys the laws of particle mechanics. This axiom can be regarded as a version of the theorem that the center of mass of a system of particles or a rigid body moves as though all the mass of system were located there, and all of the forces applied there. R2 specifies that $\theta(k, t)$ is an r by r orthogonal matrix, as is required by the intended interpretation, since the rows of $\theta(k, t)$ form a set of orthogonal unit vectors in the directions of the moving axes of k . It is necessary that this function be twice differentiable with respect to time in order that the rotational motion of the body be describable as due to finite applied torques. This axiom also rules out impacts, in which there may be discontinuous changes of angular momentum, and for which the angular acceleration does not exist.

The symmetry and positive semi-definiteness of $\mu(k)$ required by axiom R3 follows also directly from the intended interpretation of this concept (see p. 3). The restriction on the rank of the matrix $\mu(k)$ amounts to a

² The matrix cross-product AXB of two vectors A and B is defined as the difference $A*B - B*A$. This is a skew-symmetric matrix, corresponding to a symmetric double tensor. In three dimensions the matrix AXB depends only on three independent components and is closely related to the three-dimensional vector cross product representing the moment of a force B applied through a lever arm A .

restriction on the 'dimension' of the rigid body k . It can be shown that if all masses are positive, and $\mu(k)$ is defined as on page 3, then the rank of $\mu(k)$ is equal to the dimension of k , defined as the dimension of the smallest 'hyperplane' of E_r containing all of the points of k .

Axiom R4 requires that the sum $\sum_{i=1}^{\infty} H^2(k, t, i) \times H^1(k, t, i)$, which represents the resultant moment force applied to k at time t relative to the fixed coordinate axes of k , be absolutely convergent. This requirement is put on H simply in order that the resultant moment or torque should not depend on the ordering of the applied forces.

Finally, axiom R5 is a formulation of the well-known law equating rate of change of angular momentum and moment force. The matrix expression for angular momentum is $\theta(k, t) \times \left[\mu(k) \frac{d}{dt} (\theta(k, t)) \right]$, and the expression on the left side of the equation in R5 is the first time derivative of this angular momentum, which is according to this equation equal to resultant moment force.

Remark 1. The axioms for classical particle mechanics (Definition 1) do not contain any version of Newton's Third Law, nor does any version of it occur in axioms R2 to R5, and therefore our axioms for *RBM* do not include this law. This omission may seem strange in view of the fact that this is the law which justifies neglecting the internal forces acting between the parts of a rigid body in computing its motion. Two comments are in order here. First, if Newton's Third Law were not true (as it applies to internal forces within rigid bodies), it would only be necessary to represent all forces, internal as well as external, by the function H , and the equations of linear and angular acceleration (axioms P6 and R5) would still hold true. Second, the fact that the Third Law is true justifies the omission of the internal forces, and representing only the external forces by H . It would become necessary to include Newton's Third Law if a distinction between external and internal forces were made within this system, and then the force and moment force occurring in the equations of motion were defined to be the resultants of the external forces only.

Remark 2. Although Newton's Third Law is not included, our axioms satisfy two criteria of adequacy for mechanical theories. First, the well known laws of rigid motion, such as Euler's equations (Whittaker (12), p. 144), and the tensor forms, as well as the much simpler laws for two-dimensional rigid motion are derivable from our axioms. Second, it can be

shown that our equations are *deterministic* in the sense that if the initial positions and velocities of the bodies are arbitrarily prescribed, and the applied forces are given, then the paths of the bodies are uniquely determined.

Remark 3. It is to be observed that, although the primitive notions μ and θ are both defined in terms of position vectors and mass in their intended interpretations, the only *formal* connection between moment of inertia, angular position, and mass and position stated in the axioms is through axioms R5, specifying a connection with moment force, and axiom R1 (including Newton's Second Law), which in turn links resultant force with acceleration and mass. If the rotational and translational concepts were completely independent, this would have the odd consequence that it would be possible to transform the coordinate axes of the space by, say, a Galilean transformation and change the unit of mass measurement, without this being accompanied by a corresponding transformation in the amount of inertia and angular position functions. In turn, if the transformations of μ and θ were independent of those of space and mass, then the former could not be regarded as tensor quantities in the usual sense, with prescribed transformation laws. As was noted above, the translational and rotational concepts are not completely independent in this theory, since they are both linked to force. The author has not so far been able to determine, however, whether the two equations of motion place sufficient constraint on the two kinds of functions, so that the transformations of the mass and position functions uniquely determine the transformations of the moment of inertia and angular position functions.

3. Reduction. A first glance at the usual derivation of the laws of *RBM* from those of *PM* suggests that the reduction of *RBM* to *PM* consists in the following: first, the primitive notions of *RBM* are defined in terms of those of *PM*, as is indicated roughly in the intended interpretations of the primitives of *RBM*, and then the laws of *RBM* are shown to be derivable from those of *PM*, supplemented by the indicated definitions. Upon closer inspection, however, two difficulties appear in the above simple theory of reduction. The first difficulty is of a technical rather than of a conceptual nature, but is worth noting, none the less. This is simply that there are, literally, no primitive concepts in the two theories we have considered. The two theories formulated in Definitions

1 and 2 are, of course, no more than definitions, and the letters ' P ', ' T ', ' m ', ' S ', ' F ', and ' K ', ' g ', ' R ', ' H ', ' μ ', and ' θ ' are actually only variables employed in the definitions of the predicates 'system of classical r -dimensional particle mechanics,' and 'system of r -dimensional *RBM*.' Each theory, in other words, involves only one new term. This apparent difficulty is circumvented by simply replacing the definitions of the various concepts of *RBM* by a single definition which combines all of them, and which defines the predicate 'system of *RBM*' in terms of 'system of *PM*.' We shall not pursue this problem here, however, but turn our attention to the second difficulty, which is more serious.

Why, one may ask, should one bother to define the concept of a system of *RBM* in terms of that of a system of *PM*, when, in fact, both are defined in terms of the concepts of pure mathematics, as they are in Definitions 1 and 2? If it is the case that the concept of a system of *RBM* is definable in terms of set-theoretical concepts alone, as in Definition 2, and the laws of *RBM* follow from those of set theory augmented by the definition in question, then it should follow, according to the theory of reduction just proposed, that *RBM* is reducible to set theory.

On intuitive grounds, any definition of 'reduction' which has a consequence that some physical theory is reducible to set theory and analysis, seems unacceptable.

The solution we shall propose to the difficulty raised above (assuming it is felt to be one) involves a revision of the concept of a *theory* which we have been tacitly assuming up until now; i.e., that a theory — in particular the theories of *RBM* and *PM* — is simply the set-theoretical predicate defined by its axioms.^{3 4} This revision is suggested by a closer examination of the situation which prevails when one theory is reduced to another. The reduction of *RBM* to *PM* involves more than an arbitrary formal definition of the concepts of the former theory — moment of inertia and angular position — in terms of those of the latter, from which the laws of *RBM* can be shown to follow. As Nagel [8] has pointed out, these 'definitions' are actually empirical hypotheses, and as such, ones which might

³ Since the set theoretical predicate is determined by the axioms, and conversely it determines the axioms in the sense that the axioms are simply statements which are true of all and only those entities which satisfy the predicate, it makes little difference whether theories are construed as set-theoretical predicates or as sets of axioms. Thus, one would not expect to get around the difficulty by simply going over to the *linguistic* version of a theory, which construes it as a set of axioms plus all of the theorems derivable from the axioms.

⁴ See [6] for a discussion of this concept of a theory.

be false. There is, however, nothing in the account so far given of theories and their mutual relations which takes into account the fact that theories and the hypotheses represented by the 'definitions' involved in the reduction of one theory to another may be either true or false. Our first step, then, in analyzing the logic of reduction, will be to elaborate the concept of a theory in such a way that it will be possible to speak of its truth or falsity.⁵

There are undoubtedly many ways of bringing the concept of *truth* or *correctness* into formal consideration. One way, for example, is to require that the axioms be consistent with a set of observation sentences. In any case there must be some kind of reference beyond the axioms themselves to the 'things' they are supposed to describe, or to observations about those objects. We have chosen to approach this through the notion of an *intended interpretation* or an *intended model* of the theory. Very roughly speaking, an intended model of a theory is any system which, for one reason or another, it is *demand*ed that the axioms conform to. There will, in general, be a large number of systems which satisfy the axioms of a theory, but usually for theories in empirical science only a few of these will be intended applications or intended models. For example, in the case of classical *PM*, axiomatized in Definition 1, the ordered quintuple $\langle P, T, m, S, F \rangle$ such that

$$\begin{aligned} P &= \{1\} \\ T &= [0, 1] \\ m(1) &= 1 \\ S(1, t) &= \langle 0, 0, 0 \rangle \quad 0 \leq t \leq 1 \\ F(1, t, n) &= \langle 0, 0, 0 \rangle \quad 0 \leq t \leq 1; n = 1, 2, 3, \dots \end{aligned}$$

⁵ Some readers will object to speaking of the truth or falsity of a theory, and would prefer to use the terminology of confirmation. To include the concept of the confirmation of a theory relative to a given set of data would be to proceed in the same direction we propose to go: i.e., to include some connections between the fundamental or defined concepts of the theory and either observation or observation sentences, which alone will determine either the truth or the degree of confirmation of the theory. However, the theory of confirmation is at present in such an imperfect state, as it relates to theories of high complexity, that it would be extremely difficult if not impossible to found a precise analysis of reduction on it. On the other hand, the work of Tarski [11] and others on the concepts of *truth* and *satisfaction* and others relating to the interpretation of formal systems makes these concepts ideal tools for use in precise logical analyses. Our use of the concept of *truth* rather than *confirmation* is thus dictated by the requirements of logical precision; it does not imply that the author believes that in any 'ultimate' sense the concept of truth is fundamental and that of confirmation only derivative.

satisfies the axioms of PM , though it is not normally taken as an intended model simply for the reason that 1 is not a particle. On the other hand, the system in which P is the set of planets of the solar system together with the sun, m gives the masses of these objects (in some fixed units), S gives their locations relative to a system of cartesian coordinate axes fixed with respect to the fixed stars, and F gives the gravitational forces acting between the sun and planets, is an intended model of PM (or, at any rate, was often taken to be one.) It is this second kind of *intended* model which it is expected should satisfy the axioms, and the axioms or the theory is judged true or false according as the intended models satisfy the axioms or not.

If truth and falsity are to be defined, we have seen that two aspects of a theory must be brought into account: first, the formal aspect which corresponds to the set-theoretical predicate defined by the axioms (since we wish later to avoid reference to linguistic entities, such as predicates, we shall instead consider the extension of this predicate, which is the set of all systems satisfying the axioms); and second, the applied aspect, corresponding to the set of intended models. Formally, a theory T will be construed as an ordered-pair of sets $T = \langle C, I \rangle$ such that C is the set of all entities satisfying the axioms, and I is the set of intended models. We shall call C the "*characteristic set*" of T . In the case of classical PM , for example, C is the set of all ordered quintuples $\langle P, T, m, S, F \rangle$ satisfying axioms P1–P6. Just what systems are comprised within the set I of intended models for classical PM cannot be specified with precision, owing to the vagueness in the physical concepts of '*particle*,' '*position*,' '*mass*,' and '*force*.' Even to attempt an analysis of the intended models of classical PM would fall outside the scope of this paper. It will turn out, though, that such an analysis is not essential to our account of reduction, which rests on certain assumptions about the *relations* between the intended models of PM and RBM , and not on any theory as to what those models are.

One thing which it is essential to note in connection with the intended models of PM is that they are all 'physical systems' in an extended sense. They must be entities which could at least conceivably satisfy the axioms, and therefore they must be ordered-quintuples $\langle P, T, m, S, F \rangle$. Roughly, then, the intended models will be systems $\langle P, T, m, S, F \rangle$ such that P is a set of particles (physical objects whose size, for the purposes of the application, can be neglected, and not, for example, numbers), T is a set of clock readings during an interval of time, m , S , and F are functions

giving the results of measurements of mass, position, and forces applied to particles of the system during the time interval. Similarly, the intended models of *RBM* will be ordered septuples $\langle K, T, g, R, H, \mu, \theta \rangle$ satisfying the descriptions in the intended interpretations.

With theories characterized as ordered-pairs, the first member of which is its characteristic set — i.e., all entities satisfying its axioms — and the second member of which is its set of intended models, “truth” becomes definable in an obvious way. The theory is true if and only if all of its intended models satisfy its axioms, otherwise it is false. If $T = \langle C, I \rangle$, then T is true if and only if I is a subset of C .

In terms of the modified conception of theory outlined above, it is possible to give what we hope is a more adequate explication of ‘reduction’ than the one originally proposed. The ‘definition’ of the fundamental concepts of the secondary theory of the reduction (in this case *RBM*) in terms of those of the primary theory (*PM* in this case) represents, we have argued, an empirical hypothesis. This hypothesis is one which postulates that there is a certain connection between the intended models of the secondary and primary theories. In the case of *RBM* and *PM*, the assumption is that every rigid body is composed of particles, and that the masses, positions, applied forces, moments of inertia, and angular positions or the rigid bodies are related to the masses, positions, and applied forces on the particles composing them as outlined in the previous section. This assumption is clearly about the intended interpretations of *RBM* and *PM* and not about all entities satisfying their axioms, since there will be members of the characteristic set of *RBM* which are not physical objects at all, and hence not ‘composed’ of anything. Similarly, in the reduction of thermodynamics to statistical mechanics, it is assumed that all thermal bodies are composed of molecules, and that the absolute temperature of the body is proportional to the mean kinetic energy of the molecules composing it. This again is an assumption about the objects to which the two theories are applied; i.e., about their intended models. In each reduction, it is assumed that every intended model in the secondary theory has a particular relation to some intended model of the primary theory.

It is possible to formalize the above interpretation of the definition of the concepts of the secondary theory in terms of those in the primary theory as follows. Let $T_1 = \langle C_1, I_1 \rangle$ be the primary theory, and let $T_2 = \langle C_2, I_2 \rangle$ be the secondary theory which is reduced to T_1 . The ‘definition’ in question can be represented as a hypothesis that every

intended model $i_2 \in I_2$ if the secondary theory has a special relation R (which we shall call the 'reduction relation') to an intended model $i_1 \in I_1$ of the primary theory T_1 . Although we shall not attempt to formalize the informal characterizations of primary and secondary theories and reductions given above, we shall set down quasi-formally the basic connection just stated between the intended models of the primary and secondary theories and the reduction relation as Condition A, below.

CONDITION A. *Let $T_1 = \langle C_1, I_1 \rangle$ and $T_2 = \langle C_2, I_2 \rangle$ be two theories such that T_2 is reduced to T_1 by relation R . Then for all i_2 in I_2 there exists i_1 in I_1 such that $i_2 R i_1$.*

Simply defining the intended models of the secondary theory in terms of the intended models of the primary theory does not, of course, reduce one theory to the other. It must also be shown that in some sense, the laws of the secondary theory 'follow' from the laws of the primary theory together with the definition. One way to formulate this requirement, which avoids reference to such syntactical concepts as *derivability*, is as follows: it must be the case that if any element c_2 has relation R to some element c_1 which satisfies the laws of the primary theory (i.e., c_1 is in C_1), then c_2 satisfies the laws of the secondary theory (c_2 is in C_2). This second requirement is formulated explicitly in Condition B, below.

CONDITION B. *Let $T_1 = \langle C_1, I_1 \rangle$ and $T_2 = \langle C_2, I_2 \rangle$ be two theories such that T_2 is reduced to T_1 by relation R . Then for all c_1 and c_2 , if c_1 is in C_1 and $c_2 R c_1$ then c_2 is in C_2 .*

Conditions A and B do not, of course, *define* the concept of a reduction relation. However, they do have one very important consequence: if a theory T_2 is reduced to a theory T_1 by a relation R satisfying Conditions A and B, then if T_1 is correct, then T_2 is correct. Thus, any reduction relation which satisfies Conditions A and B satisfies what seems to us to be the most essential requirement for reduction, namely: it must be possible to show that if the primary theory in the reduction is correct in that all of its intended models satisfy its axioms, then all of the intended models of the secondary theory satisfy *its* axioms, and therefore the secondary theory is also correct. This is the core of the reduction of thermodynamics to statistical mechanics. In this case, what is shown is that if the laws of statistical mechanics are correct (and this may be doubtful), and the hypothesis of the reduction is correct (which says that every thermal body is composed of particles, and its temperature is

proportional to the mean kinetic energy of the particles composing it), then thermodynamics is correct.

Remark. As has been pointed out, Conditions A and B do not define the concept of reduction. A complete analysis of this notion would undoubtedly formulate considerably more restrictive conditions than ours on the concept. In fact, our conditions are so weak that for any two correct theories it is possible to construct a trivial relation 'reducing' one to the other satisfying Conditions A and B. Nagel [8] and Bergmann [2] have discussed some further restrictions informally. However, it is worth observing that conditions much like our A and B are central to both of their analyses.

4. Reduction of *RBM* to *PM*. The reduction relation relating *RBM* and *PM* can be defined by simply formalizing the descriptions of the intended interpretations of the primitive notions of *RBM* in terms of the concepts of *PM*, as outlined in Section 2. The precise formalization of this definition is too lengthy to be included here, and we shall only sketch its main features. Let *R* be the reduction relation; it is necessary to specify when *R* holds between an ordered septuple $I' = \langle K, T, g, R, H, \mu, \theta \rangle$ and an ordered quintuple $\Delta = \langle P, T, m, S, F \rangle$. In the intended interpretation of *K*, it is assumed that the elements of *K* are composed of particles — i.e., that each rigid body is a set of particles. This requirement can be formulated by imposing the condition that if *I'* has relation *R* to Δ , then *K* must be a *partition* of *P*: i.e., the particles composing *P* can be separated into sets which 'compose' the rigid bodies in *K*. In addition to the requirement that *K* be a partition of *P*, it is also necessary to impose the requirement that the particles which form a particular element of *K* maintain constant mutual distances: that is, if *p* and *q* are both elements of *k*, then for all *t* in *T*,

$$|S(p, t) - S(q, t)|$$

is a constant.

Not only must the rigid bodies *k* be composed of elements of *P*, but the mass, position, force, moment of inertia, and angular position functions of *I'* must have the proper relations to the mass, position, and force functions of Δ . For example, the mass $g(k)$ of rigid body *k* must be the sum of the masses of the particles composing it. Hence, a condition in the

definition of R must be that if Γ has relation R to \mathcal{A} , then for all k in K ,

$$g(k) = \sum_{p \in k} m(p).$$

Similarly, if $R(k, t)$ is to represent the position of the center of mass of k at time t , it must be required that for all k in K and t in T ,

$$R(k, t) = \frac{\sum_{p \in k} m(p)S(p, t)}{\sum_{p \in k} m(p)}.$$

Similar conditions relate the remaining functions H , μ and θ of Γ to the functions m , S and F of \mathcal{A} .

With the relation R defined, it is possible to ask whether or not it satisfies Conditions A and B given in Section 3. Condition A requires that every intended model of *RBM* has relation R to some intended model of *PM*. The intended models of *RBM* are systems of rigid bodies, and those of *PM* are systems of particles. That a system of rigid bodies has relation R to a system of particles means that the rigid bodies in the first system are composed of the particles in the second system, that the masses of the rigid bodies are equal to the sums of the masses of the particles composing them, and that the other functions of the rigid body system have the proper relations to the mass, position, and force functions of the particle system. Clearly whether or not Condition A is satisfied, depends on the empirical hypothesis that all rigid bodies are composed of particles which move about as though fixed in rigid frames, and the sum of whose masses is equal to the mass of the body.

The determination of whether or not Condition B is satisfied does not raise any empirical questions. In fact, it can be shown logically that if a system $\mathcal{A} = \langle P, T, m, S, F \rangle$ satisfies the axioms of *PM*, and if $\Gamma = \langle K, T, g, R, H, \mu, \theta \rangle$ has relation R to \mathcal{A} , then Γ satisfies the axioms of *RBM*. This is essentially what is proven in the usual 'derivations' of the laws of *RBM* from those of *PM* given in text books, and this is equivalent to a proof that Condition B is satisfied.

Hence it can be rigorously proven that whether relation R actually gives a reduction of *RBM* to *PM* depends on whether the empirical assumptions involved in Conditions A are correct. If they are not, then *RBM* has not been reduced to *PM*, and the usual deduction of the laws of *RBM* from those of *PM* is invalid. If, for example, there were a rigid body not composed of particles, then it is clear that nothing could be

deduced about its behavior from the laws of particle mechanics, since those laws only describe the behavior of particles.⁶

The empirical question here raised is a very difficult one, and involves in addition the problem of clarifying the rather vague notion of a *particle*. It may be observed that the molecular theory lends support to the hypothesis that rigid bodies are composed of entities small enough to approximate the point-particles required in the derivation of the laws of *RBM*, and the theory of solids indicates that these molecules remain relatively fixed within rigid bodies. However, the facts that molecules only approximate point-particles, and that they are not perfectly rigidly fixed within the bodies they compose, shows that the deduction of the laws of *RBM* from those of *PM* depends on an hypothesis which, taken exactly, is false. The necessary revisions are, however, complicated, and are, in any case, beyond the scope of this paper.

Bibliography

- [1] ADAMS, E. W., *Axiomatic Foundations of Rigid Body Mechanics*. Unpublished Ph. D. dissertation, Stanford University, 1955.
- [2] BERGMANN, G., *Philosophy of Science*. Madison 1957, XII + 181 pp.
- [3] JOOS, G., *Theoretical Physics*. Translated by I. Freeman. New York 1934, XXIII + 748 pp.
- [4] MCCONNELL, A. J., *Applications of the Absolute Differential Calculus*. London 1931, XII + 318 pp.
- [5] MCKINSEY, J. C. C., A. C. SUGAR and P. SUPPES, *Axiomatic Foundations of Classical Particle Mechanics*. Journal of Rational Mechanics, vol. 2 (1953), pp. 253–272.
- [6] MCKINSEY, J. C. C. and P. SUPPES, *Philosophy and the axiomatic foundations of physics*. Proceedings of the XIth International Congress of Philosophy, vol. VI (1953), pp. 49–53.
- [7] MILNE, E. A., *Vectorial Mechanics*. New York 1948, XII + 382 pp.
- [8] NAGEL, E., *The meaning of reduction in the natural sciences*. Reprinted in Readings in Philosophy of Science, P. P. Wiener editor, New York (1953), pp. 531–549.

⁶ It may be objected that the derivation of the laws of rigid motion includes the motions of rigid bodies with continuous mass distributions. The properties of bodies with continuous distributions are, however, derived from the laws of continuum mechanics (see, e.g., Noll, W. [9] in this volume), which is not a branch, but an extension of particle mechanics.

- [9] NOLL, W., *The foundations of classical mechanics in the light of recent advances in continuum mechanics*. This volume, pp. 226–281.
- [10] RUBIN, H. and P. SUPPES, *Transformations of systems of relativistic particle mechanics*. Pacific Journal of Mathematics, vol. 4 (1954), pp. 563–601.
- [11] TARSKI, A., *The Concept of Truth in Formalized Languages*. In Logic, Semantics, Metamathematics by A. Tarski, translated by J. H. Woodger. Oxford 1956, pp. 152–278.
- [12] WHITTAKER, E. T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed. New York 1944, XIV + 456 pp.

**THE FOUNDATIONS OF CLASSICAL MECHANICS
IN THE LIGHT OF RECENT ADVANCES
IN CONTINUUM MECHANICS¹**

WALTER NOLL

Carnegie Institute of Technology, Pittsburgh, Pennsylvania, U.S.A.

1. Introduction. It is a widespread belief even today that classical mechanics is a dead subject, that its foundations were made clear long ago, and that all that remains to be done is to solve special problems. This is not so. It is true that the mechanics of systems of a finite number of mass points has been on a sufficiently rigorous basis since Newton. Many textbooks on theoretical mechanics dismiss continuous bodies with the remark that they can be regarded as the limiting case of a particle system with an increasing number of particles. They cannot. The erroneous belief that they can had the unfortunate effect that no serious attempt was made for a long period to put classical continuum mechanics on a rigorous axiomatic basis. Only the recent advances in the theory of materials other than perfect fluids and linearly elastic solids have revived the interest in the foundations of classical mechanics. A clarification of these foundations is of importance also for the following reason. It is known that continuous matter is really made up of elementary particles. The basic laws governing the elementary particles are those of quantum mechanics. The science that provides the link between these basic laws and the laws describing the behavior of gross matter is statistical mechanics. At the present time this link is quite weak, partly because the mathematical difficulties are formidable, and partly because the basic laws themselves are not yet completely clear. A rigorous theory of continuum mechanics would give, at least some precise information on what kind of gross behavior the basic laws ought to predict.

I want to give here a brief outline of an axiomatic scheme for continuum mechanics, and I shall attempt to introduce the same level of rigor and clarity as is now customary in pure mathematics. The mathematical

¹ The results presented in this paper were obtained in the course of research sponsored by the U.S. Air Force Office of Scientific Research under contract no. AF 18 (600)-1138 with Carnegie Institute of Technology.

structures involved are quite complex, and some fine details have to be omitted in order not to overburden the paper with technicalities.

Notation: Points and vectors in Euclidean space will be indicated by bold face letters. If \mathbf{x} and \mathbf{y} are two points, then $\mathbf{x} - \mathbf{y}$ denotes the vector determined by the ordered pair (\mathbf{y}, \mathbf{x}) . If \mathbf{x} is a point and \mathbf{v} a vector, then $\mathbf{x} + \mathbf{v}$ denotes the point uniquely determined by $(\mathbf{x} + \mathbf{v}) - \mathbf{x} = \mathbf{v}$. The word "smooth" will be used instead of "continuously differentiable". Some equations will be valid only up to a set of measure zero. It will be clear from the context when this is the case.

2. Bodies.

DEFINITION 1: A BODY is a set \mathfrak{B} endowed with a structure defined by

- (a) a set Φ of mappings of \mathfrak{B} into a three-dimensional Euclidean point space E , and
- (b) a real valued set function m defined for a set of subsets of \mathfrak{B}

subject to seven axioms as follows:

- (S.1) Every mapping $\varphi \in \Phi$ is one-to-one.
- (S.2) For each $\varphi \in \Phi$, the image $B = \varphi(\mathfrak{B})$ is a region in the space E , a region being defined as a compact set with piecewise smooth boundaries.
- (S.3) If $\varphi \in \Phi$ and $\psi \in \Phi$ then the mapping $\chi = \psi \circ \varphi^{-1}$ of $\varphi(\mathfrak{B})$ onto $\psi(\mathfrak{B})$ can be extended to a smooth homeomorphism of E onto itself.
- (S.4) If χ is a smooth homeomorphism of E onto itself and if $\varphi \in \Phi$, then also $\chi \circ \varphi \in \Phi$.

These four axioms give \mathfrak{B} the structure of a piece of a differentiable manifold that is isomorphic to a region in Euclidean three-space. The following three axioms give \mathfrak{B} the structure of a measure space.

- (M.1) m is a non-negative measure, defined for all Borel subsets \mathfrak{C} of \mathfrak{B} .
- (M.2) For each $\varphi \in \Phi$, the measure $\mu_\varphi = m \circ \varphi^{-1}$ induced by m on the region $B = \varphi(\mathfrak{B})$ in space is absolutely continuous relative to the Lebesgue measure in B . Hence it has a density ρ_φ so that

$$(2.1) \quad m(\mathfrak{C}) = \int_{\varphi(\mathfrak{C})} \rho_\varphi(\mathbf{x}) dV.$$

- (M.3) For each $\varphi \in \Phi$ the density ρ_φ is positive and bounded.

² The symbol \circ denotes the composition of mappings and a superposed -1 denotes the inverse of a mapping.

We use the following terminology: The elements X, Y, \dots of \mathfrak{B} are the PARTICLES of the body. The mappings $\varphi \in \Phi$ are the CONFIGURATIONS of the body. The point $\mathfrak{x} = \varphi(X)$ is the POSITION of the particle X in the configuration φ . The set function m is the MASS DISTRIBUTION of the body. The number $m(\mathfrak{C})$ is the MASS of the set \mathfrak{C} . Here and subsequently we refer to Borel sets simply as sets. The density ρ_φ is the MASS DENSITY of \mathfrak{B} in the configuration φ . Note that it would have been sufficient to require the existence of ρ_φ only for one particular configuration φ . It then follows that the mass density exists also for all other configurations.

A compact subset \mathfrak{P} of \mathfrak{B} with piecewise smooth boundaries will be called a PART of the body \mathfrak{B} . It may again be regarded as a body whose configurations are the restrictions to \mathfrak{P} of the configurations of \mathfrak{B} and whose mass distribution is the restriction of the mass distribution of \mathfrak{B} to the subsets of \mathfrak{P} . Two parts \mathfrak{P} and \mathfrak{Q} will be called SEPARATE if

$$\mathfrak{P} \cap \mathfrak{Q} \subset \overline{\mathfrak{P}} \cap \overline{\mathfrak{Q}},$$

where $\overline{\mathfrak{P}}$ denotes the boundary of \mathfrak{P} .

3. Kinematics

DEFINITION 2: A MOTION of a body \mathfrak{B} is a one-parameter family $\{\theta_t\}$, $-\infty < t < \infty$, of configurations $\theta_t \in \Phi$ of \mathfrak{B} such that

(K.1) *The derivative*

$$(3.1) \quad \mathbf{v}(X, t) = \frac{d}{dt} \theta_t(X)$$

exists for all $X \in \mathfrak{B}$ and all t , it is a continuous function of X and t jointly, and it is a smooth function of X .

(K.2) *The derivative*

$$(3.2) \quad \dot{\mathbf{v}}(X, t) = \frac{d}{dt} \mathbf{v}(X, t) = \frac{d^2}{dt^2} \theta_t(X)$$

exists piecewise and is piecewise continuous in X and t jointly.

The parameter t is called the TIME. Derivatives with respect to t will be denoted by superposed dots. $\mathbf{v}(X, t)$ is called the VELOCITY of the particle X at time t . $\dot{\mathbf{v}}(X, t)$ is called the ACCELERATION of X at t .

Let h be any real, vector, or tensor valued function of X and t , and assume that $h(X, t)$ is smooth in X and t jointly. We may then associate

with h the function \hat{h} defined by

$$(3.3) \quad \hat{h}(\mathbf{x}, t) = h(\theta_t^{-1}(\mathbf{x}), t)$$

for $-\infty < t < \infty$ and $\mathbf{x} \in \theta_t(\mathfrak{B})$. By the chain rule of differentiation we have

$$(3.4) \quad h(X, t) = \dot{\hat{h}}(\theta_t(X), t) + \nabla \hat{h}(\theta_t(X), t) \cdot \mathbf{v}(X, t),$$

where $\nabla \hat{h}$ denotes the gradient of \hat{h} with respect to \mathbf{x} . It is customary in the literature to use the same symbol for h and \hat{h} , to omit the independent variables, and to distinguish $\dot{\hat{h}}$ from \hat{h} by writing $\dot{\hat{h}} = \frac{\partial \hat{h}}{\partial t}$. Equation (3.4) then takes the familiar form

$$(3.5) \quad h = \frac{\partial h}{\partial t} + \mathbf{v} \cdot \text{grad } h.$$

The LINEAR MOMENTUM at time t of a set $\mathfrak{C} \subset \mathfrak{B}$ is defined by

$$(3.6) \quad \mathbf{g}(\mathfrak{C}; t) = \int_{\mathfrak{C}} \mathbf{v}(X, t) dm.$$

It follows from (K.1) and (K.2) that $\mathbf{g}(\mathfrak{C}, t)$ is piecewise smooth in t . As a function of \mathfrak{C} it is a vector valued measure, absolutely continuous relative to m with density \mathbf{v} .

The ANGULAR MOMENTUM at time t of a set $\mathfrak{C} \subset \mathfrak{B}$, relative to a point $\mathbf{O} \in E$, is defined by

$$(3.7) \quad \mathbf{h}(\mathfrak{C}; t; \mathbf{O}) = \int_{\mathfrak{C}} [\theta_t(X) - \mathbf{O}] \times \mathbf{v}(X, t) dm.$$

It is piecewise smooth in t , and, as a function of \mathfrak{C} , it is a vector valued measure.

4. Forces

DEFINITION 3: A SYSTEM OF BODY FORCES for a body \mathfrak{B} is a family $\{\mathbf{B}_{\mathfrak{P}}\}$ of vector valued set functions subject to the following axioms:

- (B.1) For each part \mathfrak{P} of \mathfrak{B} , $\mathbf{B}_{\mathfrak{P}}$ is a vector valued measure defined on the Borel subsets of \mathfrak{P} .
- (B.2) For each \mathfrak{P} , $\mathbf{B}_{\mathfrak{P}}$ is absolutely continuous relative to the mass distribution m of \mathfrak{P} . Hence it has a density $\mathbf{b}_{\mathfrak{P}}$ so that

$$(4.1) \quad \mathbf{B}_{\mathfrak{P}}(\mathfrak{C}) = \int_{\mathfrak{C}} \mathbf{b}_{\mathfrak{P}}(X) dm.$$

(B.3) The density $\mathbf{b}_{\mathfrak{P}}$ is bounded, i.e.

$$|\mathbf{b}_{\mathfrak{P}}(X)| < k < \infty,$$

where k is independent of \mathfrak{P} and $X \in \mathfrak{P}$.

DEFINITION 4: A SYSTEM OF CONTACT FORCES for a body \mathfrak{B} is a family $\{\mathbf{C}_{\mathfrak{P}}\}$ of vector valued set functions subject to the following axioms:

(C.1) For each part \mathfrak{P} of \mathfrak{B} , $\mathbf{C}_{\mathfrak{P}}$ is a vector valued measure defined on the Borel subsets of \mathfrak{P} .

(C.2) $\mathbf{C}_{\mathfrak{P}}(\mathfrak{C}) = \mathbf{C}_{\mathfrak{P}}(\mathfrak{C} \cap \overline{\mathfrak{P}})$.

(C.3) If $\mathfrak{c} \subset \overline{\mathfrak{D}}$, $\mathfrak{c} \subset \overline{\mathfrak{P}}$, and $\mathfrak{P} \subset \mathfrak{D}$, then

$$\mathbf{C}_{\mathfrak{P}}(\mathfrak{c}) = \mathbf{C}_{\mathfrak{D}}(\mathfrak{c}).$$

(C.4) If $\varphi \in \Phi$ is any configuration of \mathfrak{B} and if $\overline{P} = \varphi(\overline{\mathfrak{P}})$, then the induced measure $\mathbf{C}_{\mathfrak{P}} \circ \varphi^{-1}$, when restricted to the Borel subsets of the boundary surface \overline{P} of $P = \varphi(\mathfrak{P})$, is absolutely continuous relative to the Lebesgue surface measure on \overline{P} . Hence it has a density $\mathbf{s}(\mathfrak{P}, \varphi)$ so that

$$(4.2) \quad \mathbf{C}_{\mathfrak{P}}(\mathfrak{c}) = \int_{\varphi(\mathfrak{c})} \mathbf{s}(\mathfrak{P}, \varphi; \mathbf{x}) dA$$

for all Borel subsets $\mathfrak{c} \subset \overline{\mathfrak{P}}$.

(C.5) The density $\mathbf{s}(\mathfrak{P}, \varphi)$ is bounded, i.e.

$$|\mathbf{s}(\mathfrak{P}, \varphi; \mathbf{x})| < l < \infty,$$

where l does not depend on \mathfrak{P} or $\mathbf{x} \in \varphi(\overline{\mathfrak{P}})$.

As in the case of a mass distribution, it would, have been sufficient in (C.4) to require the existence of $\mathbf{s}(\mathfrak{P}, \varphi)$ only for a particular $\varphi \in \Phi$. The existence of \mathbf{s} for all other configurations is then an automatic consequence. The axiom (C.2) means that $\mathbf{C}_{\mathfrak{P}}$ is essentially a vector measure on the boundary $\overline{\mathfrak{P}}$.

It is useful to consider surfaces in \mathfrak{B} as being *oriented*, and to employ the operation of addition of oriented surfaces in the sense of algebraic topology. The boundary $\overline{\mathfrak{P}}$ of a part \mathfrak{P} of \mathfrak{B} will be regarded as oriented in such a way that the positive side of $\overline{\mathfrak{P}}$ is exterior to \mathfrak{P} . If \mathfrak{P} and \mathfrak{Q} are two separate parts of \mathfrak{B} , then

$$(4.3) \quad \overline{\mathfrak{P} \cup \mathfrak{Q}} = \overline{\mathfrak{P}} + \overline{\mathfrak{Q}}.$$

This is true because the common boundary of \mathfrak{P} and \mathfrak{Q} , if any, appears

twice with opposite orientation on the right side of (4.3) and hence cancels. We shall say that the surface \mathfrak{c} is a *PIECE* of the surface \mathfrak{d} if \mathfrak{c} is a subset of \mathfrak{d} and if the orientation of \mathfrak{c} is induced by \mathfrak{d} . The significance of the axiom (C.3) is brought out by the following theorem:

THEOREM I: *There is a vector valued function \mathbf{S} , defined for all oriented surfaces \mathfrak{c} in \mathfrak{B} , such that*

$$(4.4) \quad \mathbf{C}_{\mathfrak{B}}(\mathfrak{c}) = \mathbf{S}(\mathfrak{c})$$

whenever \mathfrak{c} is a piece of the boundary $\overline{\mathfrak{B}}$ of \mathfrak{B} . We say that $\mathbf{S}(\mathfrak{c})$ is the CONTACT FORCE ACTING ACROSS THE ORIENTED SURFACE \mathfrak{c} .

Proof: For each \mathfrak{c} which is not a piece of $-\overline{\mathfrak{B}}$ we can find a part $\mathfrak{Q}(\mathfrak{c})$ of \mathfrak{B} such that \mathfrak{c} is a piece of $\overline{\mathfrak{Q}(\mathfrak{c})}$. We then define $\mathbf{S}(\mathfrak{c}) = \mathbf{C}_{\mathfrak{Q}(\mathfrak{c})}(\mathfrak{c})$. Now let \mathfrak{P} be an arbitrary part of \mathfrak{B} and let \mathfrak{c} be a piece of $\overline{\mathfrak{P}}$. We then have

$$\begin{aligned} \mathfrak{c} \subset \overline{\mathfrak{P}}, \quad \mathfrak{c} \subset \overline{\mathfrak{Q}(\mathfrak{c})}, \quad \mathfrak{c} \subset \overline{\mathfrak{Q}(\mathfrak{c}) \cap \mathfrak{P}}, \\ \mathfrak{P} \cap \mathfrak{Q}(\mathfrak{c}) \subset \mathfrak{P}, \quad \mathfrak{P} \cap \mathfrak{Q}(\mathfrak{c}) \subset \mathfrak{Q}(\mathfrak{c}). \end{aligned}$$

Applying axiom (C.3) twice, we get

$$\mathbf{C}_{\mathfrak{P}}(\mathfrak{c}) = \mathbf{C}_{\mathfrak{P} \cap \mathfrak{Q}(\mathfrak{c})}(\mathfrak{c}), \quad \mathbf{C}_{\mathfrak{Q}(\mathfrak{c})}(\mathfrak{c}) = \mathbf{C}_{\mathfrak{P} \cap \mathfrak{Q}(\mathfrak{c})}(\mathfrak{c}).$$

Hence

$$\mathbf{C}_{\mathfrak{P}}(\mathfrak{c}) = \mathbf{C}_{\mathfrak{Q}(\mathfrak{c})}(\mathfrak{c}) = \mathbf{S}(\mathfrak{c}).$$

If \mathfrak{c} is a part of $-\overline{\mathfrak{B}}$ we define

$$(4.5) \quad \mathbf{S}(\mathfrak{c}) = -\mathbf{S}(-\mathfrak{c}).$$

It follows from theorem I and axiom (C.4) that there is a vector valued function $\mathbf{s}(\mathfrak{c}, \varphi; \mathbf{x})$ such that

$$(4.6) \quad \mathbf{S}(\mathfrak{c}) = \int_{\varphi(\mathfrak{c})} \mathbf{s}(\mathfrak{c}, \varphi; \mathbf{x}) dA.$$

Also, if $\mathbf{x} \in \varphi(\mathfrak{d}) \subset \varphi(\mathfrak{c})$ and if \mathfrak{d} is a piece of \mathfrak{c} , then

$$(4.7) \quad \mathbf{s}(\mathfrak{c}, \varphi; \mathbf{x}) = \mathbf{s}(\mathfrak{d}, \varphi; \mathbf{x}).$$

If \mathfrak{c}_1 and \mathfrak{c}_2 are two pieces of a surface \mathfrak{c} and if $\mathfrak{c} = \mathfrak{c}_1 + \mathfrak{c}_2$, then

$$(4.8) \quad \mathbf{S}(\mathfrak{c}) = \mathbf{S}(\mathfrak{c}_1) + \mathbf{S}(\mathfrak{c}_2).$$

This is true because $\mathbf{C}_{\mathfrak{B}}$, as a measure, is additive and because, by axiom (C.4) the value of $\mathbf{C}_{\mathfrak{B}}$ for the common boundary curve of \mathfrak{c}_1 and \mathfrak{c}_2 is zero.

DEFINITION 5: A SYSTEM OF FORCES for a body \mathfrak{B} is a family of vector valued measures $\{F_{\mathfrak{P}}\}$ such that, for each part \mathfrak{P} of \mathfrak{B} , $F_{\mathfrak{P}}$ is defined on the subsets of \mathfrak{P} and such that the $F_{\mathfrak{P}}$ have decompositions

$$(4.9) \quad F_{\mathfrak{P}} = B_{\mathfrak{P}} + C_{\mathfrak{P}},$$

where $\{B_{\mathfrak{P}}\}$ is a system of body forces and $\{C_{\mathfrak{P}}\}$ is a system of contact forces.

It is not hard to see that the decomposition (4.9), if it exists, is automatically unique.

We use the following terminology: The measure $F_{\mathfrak{P}}$ is the FORCE acting on the part \mathfrak{P} of \mathfrak{B} . The vector $F_{\mathfrak{P}}(\mathfrak{P})$ is the RESULTANT FORCE acting on \mathfrak{P} . Let \mathfrak{P} and \mathfrak{Q} be two separate parts of \mathfrak{B} . The vector measure

$$(4.10) \quad F_{\mathfrak{P},\mathfrak{Q}} = F_{\mathfrak{P}} - F_{\mathfrak{P} \cup \mathfrak{Q}}$$

defined on the subsets of \mathfrak{P} , is the MUTUAL FORCE exerted on \mathfrak{P} by \mathfrak{Q} . The mutual force exerted on a part \mathfrak{P} of \mathfrak{B} by the closure of its complement is denoted by $F_{(\mathfrak{P})}$ and it is called the INTERNAL FORCE acting on \mathfrak{P} . The restriction of $F_{\mathfrak{P}}$ to a part \mathfrak{P} of \mathfrak{B} is the EXTERNAL FORCE acting on \mathfrak{P} . A similar terminology and notation will be used when "force" is replaced by "body force" or by "contact force".

Let $\{F_{\mathfrak{P}}\}$ be a system of forces for a body \mathfrak{B} , $\varphi \in \Phi$ a configuration of \mathfrak{B} , and $\mathbf{O} \in E$ a point in space. The MOMENT about \mathbf{O} of the force $F_{\mathfrak{P}}$ acting on the part \mathfrak{P} of \mathfrak{B} in the configuration φ is the vector valued measure $M(F_{\mathfrak{P}}, \varphi, \mathbf{O})$ defined by

$$(4.11) \quad M(F_{\mathfrak{P}}, \varphi, \mathbf{O}; \mathfrak{C}) = \int_{\mathfrak{C}} [\varphi(X) - \mathbf{O}] \times dF_{\mathfrak{P}}$$

for the subsets \mathfrak{C} of \mathfrak{P} . The vector $M(F_{\mathfrak{P}}, \varphi, \mathbf{O}; \mathfrak{P})$ is the RESULTANT MOMENT about \mathbf{O} acting on \mathfrak{P} .

5. Dynamical processes

DEFINITION 6: A DYNAMICAL PROCESS is a triple $\{\mathfrak{B}, \theta_t, F_{\mathfrak{P},t}\}$, where \mathfrak{B} is a body, θ_t is a motion of \mathfrak{B} , and $F_{\mathfrak{P},t}$ is a one-parameter family of systems of forces for \mathfrak{B} , subject to the following two axioms:

(D.1) *Principle of linear momentum:* For all parts \mathfrak{P} of \mathfrak{B} and all times t ,

$$(5.1) \quad F_{\mathfrak{P},t}(\mathfrak{P}) = \dot{\mathbf{g}}(\mathfrak{P}; t),$$

where \mathbf{g} is defined by (3.6). In words: The resultant force acting on the part \mathfrak{P} is equal to the rate of change of the linear momentum of \mathfrak{P} .

(D.2) *Principle of angular momentum:* Let $\mathbf{O} \in E$ be any point in space.

Then for all parts \mathfrak{P} of \mathfrak{B} and all times t ,

$$(5.2) \quad \mathbf{M}(\mathbf{F}_{\mathfrak{P},t}, \theta_t, \mathbf{O}; \mathfrak{P}) = \dot{\mathbf{h}}(\mathfrak{P}; t; \mathbf{O}),$$

where \mathbf{h} and \mathbf{M} are defined by (3.7) and (4.11), respectively. In words: The resultant moment about \mathbf{O} acting on a part \mathfrak{P} is equal to the rate of change of the angular momentum of \mathfrak{P} relative to \mathbf{O} .

It would have been sufficient to require that (5.2) holds for a particular $\mathbf{O} \in E$. It is then automatically valid for all points in space. Also, (5.2) remains valid if the fixed point \mathbf{O} is replaced by the variable mass center

$$(5.3) \quad \mathbf{c}(\mathfrak{P}, t) = \mathbf{O} + \frac{1}{m(\mathfrak{P})} \int_{\mathfrak{P}} (\theta_t(X) - \mathbf{O}) dm$$

of the part \mathfrak{P} . These statements can be proved in the classical manner.

We now prove a number of important theorems. For simplicity we omit the variable t ; we write

$$(5.4) \quad \mathbf{s}(\mathbf{c}; \mathbf{x}) = \mathbf{s}(\mathbf{c}, \theta_t; \mathbf{x})$$

for the density of the contact force as defined by (4.6).

THEOREM II: For any two separate parts \mathfrak{P} and \mathfrak{Q} of \mathfrak{B} we have

$$(5.5) \quad \mathbf{F}_{\mathfrak{P}, \mathfrak{Q}}(\mathfrak{P}) = -\mathbf{F}_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{Q})$$

i.e. the resultant mutual force exerted on \mathfrak{P} by \mathfrak{Q} is equal and opposite to the resultant mutual force exerted on \mathfrak{Q} by \mathfrak{P} .

Proof: We apply axiom (D.1) to \mathfrak{P} , \mathfrak{Q} , and $\mathfrak{P} \cup \mathfrak{Q}$:

$$(5.6) \quad \mathbf{F}_{\mathfrak{P}}(\mathfrak{P}) = \dot{\mathbf{g}}(\mathfrak{P}), \quad \mathbf{F}_{\mathfrak{Q}}(\mathfrak{Q}) = \dot{\mathbf{g}}(\mathfrak{Q}), \quad \mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P} \cup \mathfrak{Q}) = \dot{\mathbf{g}}(\mathfrak{P} \cup \mathfrak{Q}).$$

Since $\mathfrak{P} \cap \mathfrak{Q}$ has no mass by (M.2), it follows from (3.6) that

$$\mathbf{g}(\mathfrak{P} \cup \mathfrak{Q}) = \mathbf{g}(\mathfrak{P}) + \mathbf{g}(\mathfrak{Q});$$

hence, by (5.6),

$$\mathbf{F}_{\mathfrak{P}}(\mathfrak{P}) + \mathbf{F}_{\mathfrak{Q}}(\mathfrak{Q}) = \mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P} \cup \mathfrak{Q}).$$

It is not hard to see that $\mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P} \cap \mathfrak{Q}) = 0$. Hence

$$\mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P} \cup \mathfrak{Q}) = \mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P}) + \mathbf{F}_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{Q}).$$

The assertion follows now from the definition (4.10).

THEOREM III (reaction principle)³: The contact force $S(c)$ acting cross c is opposite to the contact force acting across $-c$, i.e.

$$(5.7) \quad S(c) = -S(-c)$$

Proof: If c is a piece of $-\bar{\mathfrak{B}}$, then (5.7) is true by the definition (4.5). If not, it is possible to find two separate parts \mathfrak{P} and \mathfrak{Q} such that $\mathfrak{P} \cap \mathfrak{Q} = c$ (see Fig. 1). We orient c such that it is a piece of $\bar{\mathfrak{P}}$. Then $-c$

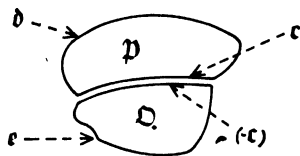


Fig. 1

will be a piece of $\bar{\mathfrak{Q}}$. The surfaces $\bar{\mathfrak{P}}$, $\bar{\mathfrak{Q}}$, and $\overline{\mathfrak{P} \cup \mathfrak{Q}}$ have decompositions

$$\bar{\mathfrak{P}} = c + b, \quad \bar{\mathfrak{Q}} = (-c) + e, \quad \overline{\mathfrak{P} \cup \mathfrak{Q}} = b + e.$$

It follows from theorem I and (4.8) that

$$C_{\mathfrak{P}}(\mathfrak{P}) = S(c) + S(b), \quad C_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P}) = S(b)$$

and hence that

$$C_{\mathfrak{P}, \mathfrak{Q}}(\mathfrak{P}) = C_{\mathfrak{P}}(\mathfrak{P}) - C_{\mathfrak{P} \cup \mathfrak{Q}}(\mathfrak{P}) = S(c).$$

Similarly, we obtain

$$C_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{Q}) = S(-c).$$

For the total resultant mutual forces, we get

$$(5.8) \quad \begin{aligned} F_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{P}) &= B_{\mathfrak{P}, \mathfrak{Q}}(\mathfrak{P}) + S(c) \\ F_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{Q}) &= B_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{Q}) + S(-c). \end{aligned}$$

Application of theorem II gives

$$(5.9) \quad S(c) + S(-c) = -[B_{\mathfrak{P}, \mathfrak{Q}}(\mathfrak{P}) + B_{\mathfrak{Q}, \mathfrak{P}}(\mathfrak{Q})].$$

Using axiom (M.3) one can show that the parts \mathfrak{P} and \mathfrak{Q} can be chosen

³ Various statements, mostly quite vague, pass under the title "principle of action and reaction" in the literature. All of these statements, when made precise, are provable theorems in the theory presented here.

such that their masses $m(\mathfrak{B})$ and $m(\mathfrak{D})$ are arbitrarily small. Axiom (B.3) then implies that the right side of (5.9) can be made arbitrarily small in absolute value. It follows that the left side of (5.9) must vanish. Q.e.d.

As a corollary, it follows that

$$(5.10) \quad \mathcal{S}(c_1 + c_2) = \mathcal{S}(c_1) + \mathcal{S}(c_2),$$

no matter whether c_1 and c_2 are pieces of $c = c_1 + c_2$, as in (4.8), or not. Hence \mathcal{S} may be regarded as an additive vector valued function of oriented surfaces in \mathfrak{B} . Another corollary is that the statement of theorem II remains true if "mutual force" there is replaced by "mutual contact force" or by "mutual body force".

THEOREM IV (stress principle)⁴: *There is a vector valued function $\mathbf{s}(\mathbf{x}, \mathbf{n})$, where $\mathbf{x} \in \theta_t(\mathfrak{B})$ and where \mathbf{n} is a unit vector, such that*

$$(5.11) \quad \mathbf{s}(c; \mathbf{x}) = \mathbf{s}(\mathbf{x}, \mathbf{n})$$

whenever $\theta_t(c)$ has the unit normal \mathbf{n} at $\mathbf{x} \in \theta_t(c)$, directed towards the positive side of the oriented surface $\theta_t(c)$, the orientation of $\theta_t(c)$ being induced by the orientation of c .

Proof: Let c_1 and c_2 be two surfaces in \mathfrak{B} tangent to each other at $\mathbf{x} = \theta_t^{-1}(\mathbf{x})$. The surfaces $c_1 = \theta_t(c_1)$ and $c_2 = \theta_t(c_2)$ in space E are then tangent to each other at the point \mathbf{x} . We assume that \mathbf{n} is their unit normal at \mathbf{x} and that c_1 and c_2 are oriented in such a way that \mathbf{n} is directed

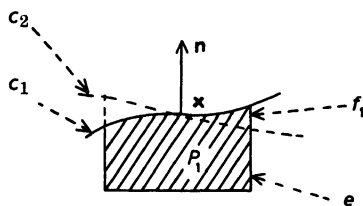


Fig. 2

toward the positive side of c_1 and c_2 . Consider the region P_1 bounded by a piece d_1 of c_1 , a piece of a circular cylinder f of radius r whose axis is \mathbf{n}

⁴ The assertion of this theorem appears in all of the past literature as an assumption. It has been proposed occasionally that one should weaken this assumption and allow the stress to depend not only on the tangent plane at \mathbf{x} , but also on the curvature of the surface c at \mathbf{x} . The theorem given here shows that such dependence on the curvature, or on any other local property of the surface at \mathbf{x} , is impossible.

and by a plane perpendicular to \mathbf{n} at a distance r from \mathbf{x} as shown in Fig. 2. The region P_2 is defined in a similar manner. Denote the common boundary of P_1 and P_2 on the cylinder and the plane by e . The boundaries \bar{P}_1 and \bar{P}_2 then have decompositions into separate pieces of the form

$$(5.12) \quad \bar{P}_1 = d_1 + e + f_1, \quad \bar{P}_2 = d_2 + e + f_2,$$

where f_1 and f_2 are pieces of the cylinder f . We denote the surface area of a surface c by $A(c)$ and the volume of a region P by $V(P)$. It is not hard to see that then

$$(5.13) \quad A(d_i) = \pi r^2 + o(r^2),$$

$$(5.14) \quad A(f_i) = o(r^2),$$

$$(5.15) \quad V(P_i) = o(r^2)$$

for $i = 1, 2$. $\mathfrak{P}_1 = \bar{\theta}_t(P_1)$ and $\mathfrak{P}_2 = \bar{\theta}_t(P_2)$ will be parts of \mathfrak{B} for small enough r , except when $\mathbf{x} \in \theta_t(\mathfrak{B})$, and \mathbf{n} is directed toward the interior of $\theta_t(\mathfrak{B})$. Applying axiom (D.1) to \mathfrak{P}_1 and \mathfrak{P}_2 gives

$$(5.16) \quad \mathbf{F}_{\mathfrak{P}_i}(\mathfrak{P}_i) = \mathbf{B}_{\mathfrak{P}_i}(\mathfrak{P}_i) + \mathbf{C}_{\mathfrak{P}_i}(\mathfrak{P}_i) = \int_{\mathfrak{P}_i} \dot{\mathbf{v}} dm, \quad i = 1, 2.$$

By (4.1) and (4.4) this may be written in the form

$$(5.17) \quad \mathbf{S}(\bar{\mathfrak{P}}_i) = \int_{\mathfrak{P}_i} (\dot{\mathbf{v}} - \mathbf{b}_{\mathfrak{P}_i}) dm.$$

By (4.6), (4.7), (4.8), and (5.12) we have

$$(5.18) \quad \mathbf{S}(\bar{\mathfrak{P}}_i) = \int_{d_i} \mathbf{s}(c_i) dA + \int_{f_i} \mathbf{s}(f_i) dA + \int_e \mathbf{s}(e) dA; \quad i = 1, 2,$$

where $\bar{f}_i = \bar{\theta}_t(f_i)$, $\bar{c} = \bar{\theta}_t(e)$. Subtracting the two equations (5.18) and using (5.17), we get

$$(5.19) \quad \int_{d_1} \mathbf{s}(c_1) dA - \int_{d_2} \mathbf{s}(c_2) dA = \\ = \int_{\mathfrak{P}_1} (\dot{\mathbf{v}} - \mathbf{b}_{\mathfrak{P}_1}) dm - \int_{\mathfrak{P}_2} (\dot{\mathbf{v}} - \mathbf{b}_{\mathfrak{P}_2}) dm - \int_{f_1} \mathbf{s}(f_1) dA + \int_{f_2} \mathbf{s}(f_2) dA.$$

Since $\dot{\mathbf{v}}$, $\mathbf{b}_{\mathfrak{P}_i}$ and the mass density are bounded by constants independent of \mathfrak{P} , according to the axioms (K.2), (B.3), and (M.3), it follows from (5.15) that

$$\int_{\mathfrak{P}_i} (\dot{\mathbf{v}} - \mathbf{b}_{\mathfrak{P}_i}) dm = o(r^2), \quad i = 1, 2.$$

Similarly, it follows from axiom (C.5) and from (5.14) that

$$\int_{f_i} \mathbf{s}(\mathbf{f}_i) dA = o(r^2), \quad i = 1, 2.$$

Hence, by (5.19),

$$\int_{d_1} \mathbf{s}(\mathbf{c}) dA = \int_{d_2} \mathbf{s}(\mathbf{c}_2) dA + o(r^2).$$

Dividing by πr^2 and using (5.13), we get

$$(5.20) \quad \frac{\int_{d_1} \mathbf{s}(\mathbf{c}_1) dA}{A(d_1)} = \frac{\int_{d_2} \mathbf{s}(\mathbf{c}_2) dA}{A(d_2)} + \frac{o(r^2)}{\pi r^2}.$$

By a theorem on measures with density, we have

$$\lim_{r \rightarrow 0} \frac{\int_{d_i} \mathbf{s}(\mathbf{c}_i) dA}{A(d_i)} = \mathbf{s}(\mathbf{c}_i; \mathbf{x}), \quad i = 1, 2.$$

Thus, letting r go to zero in (5.20), we finally obtain

$$\mathbf{s}(\mathbf{c}_1; \mathbf{x}) = \mathbf{s}(\mathbf{c}_2; \mathbf{x}),$$

which shows that $\mathbf{s}(\mathbf{c}; \mathbf{x})$ has the same value for all surfaces \mathbf{c} with the same unit normal \mathbf{n} . The exceptional case when $\mathbf{x} \in \overline{\theta_t(\mathfrak{B})}^{-1}$ and \mathbf{n} is directed toward the interior of $\theta_t(\mathfrak{B})$ is taken care of by the definition (4.5).

The vector $\mathbf{s}(\mathbf{x}, \mathbf{n})$ is called the **STRESS** acting at \mathbf{x} across the surface element with unit normal \mathbf{n} . By (4.6) the contact force $\mathbf{S}(\mathbf{c})$ acting across \mathbf{c} is given by

$$(5.21) \quad \mathbf{S}(\mathbf{c}) = \int_{\theta_t(\mathbf{c})} \mathbf{s}(\mathbf{x}, \mathbf{n}) dA,$$

where \mathbf{n} is the unit normal at \mathbf{x} to the oriented surface $\theta_t(\mathbf{c})$. By theorem II we have

$$(5.22) \quad \mathbf{s}(\mathbf{x}, \mathbf{n}) = -\mathbf{s}(\mathbf{x}, -\mathbf{n}).$$

The following two additional assumptions suffice to ensure the validity of the classical theorems of continuum mechanics:

- (a) The stress $\mathbf{s}(\mathbf{x}, \mathbf{n})$, for each \mathbf{n} , is a smooth function of $\mathbf{x} \in \theta_t(\mathfrak{B})$.

(b) For almost all $X \in \mathfrak{B}$, the limit

$$(5.23) \quad \mathbf{b}(X) = \lim_{\mathfrak{B} \rightarrow X} \frac{1}{m(\mathfrak{B})} \mathbf{B}_{\mathfrak{B}}(\mathfrak{B}),$$

where \mathfrak{B} is a neighborhood of X contracting to X , exists.

Under these assumptions, one can prove the following theorems in the classical manner:

(1) There is a field of linear transformations $S(\mathbf{x})$, $\mathbf{x} \in \theta_t(\mathfrak{B})$, such that

$$(5.24) \quad \mathbf{s}(\mathbf{x}, \mathbf{n}) = S(\mathbf{x})\mathbf{n}.$$

$S(\mathbf{x})$ is called the **STRESS TENSOR** at \mathbf{x} .

(2) The stress tensor $S(\mathbf{x})$ is symmetric.

(3) Cauchy's equation of motion

$$(5.25) \quad \operatorname{div} S + \rho \mathbf{b} = \rho \dot{\mathbf{v}}$$

holds, where S is the stress tensor, ρ is the mass density, $\dot{\mathbf{v}}$ is the acceleration, and \mathbf{b} is defined by (5.23).

6. Equivalence of dynamical processes. The position of a particle can be specified physically not in an absolute sense but only relative to a given *frame of reference*. Such a frame is a set of objects whose mutual distances change very little in time, like the walls of a laboratory, the fixed stars, or the wooden horses on a merry-go-round. In classical physics, a change of frame corresponds to a transformation of space and time which preserves distances and time intervals. It is well known that the most general such transformation is of the form

$$(6.1) \quad \begin{aligned} \mathbf{x}^* &= \mathbf{c}(t) + Q(t)(\mathbf{x} - \mathbf{O}), \\ t^* &= t + a, \end{aligned}$$

where $\mathbf{c}(t)$ is a point valued function of t , $Q(t)$ is a function of t whose values are orthogonal transformations, a is a real constant, and \mathbf{O} is a point, which may be fixed once and for all. We assume that $\mathbf{c}(t)$ and $Q(t)$ are twice continuously differentiable. A change of frame (6.1) also induces a transformation on vectors and tensors. A vector \mathbf{u} , for example, is transformed into

$$(6.2) \quad \mathbf{u}^* = Q(t)\mathbf{u}.$$

Let $\{\mathfrak{B}, \theta_t, \mathbf{F}_{\mathfrak{B}, t}\}$ be a dynamical process. A change of frame $\{\mathbf{c}, Q, a\}$

will transform the motion θ_t into a new motion θ'_t defined by

$$(6.3) \quad \theta'_t(X) = \mathbf{c}(t - a) + Q(t - a)[\theta_{t-a}(X) - \mathbf{O}].$$

The velocities and the accelerations of the two motions θ_t and θ'_t are, in general, not related by the transformation formula (6.2) for vectors. They depend on the choice of the frame of reference. We say that they are not *objective*. However, there are objective kinematical quantities, for example the rate of deformation tensor.

If we wish to assume that forces have an objective meaning we would have to require that $\mathbf{F}_{\mathfrak{B},t}(\mathfrak{C})$ transforms according to the law (6.2) under a change of frame. However, when this assumption is made, a dynamical process does not transform into a dynamical process because the axioms (D.1) and (D.2) are not preserved, except when \mathbf{c} is linear in t and Q is constant. It is this difficulty which has led to the concept of absolute space and which has caused much controversy in the history of mechanics. A clarification was finally given by Einstein in his general theory of relativity, in which gravitational forces and inertial forces cannot be separated from each other in an objective manner. If we wish to stay in the realm of classical mechanics we may resolve the paradox by sacrificing the objectivity of the external body forces while retaining the objectivity of the essential types of forces, the contact forces and the mutual body forces. This can be done by assuming that the forces transform according to a law of the form

$$(6.4) \quad \mathbf{F}'_{\mathfrak{B},t}(\mathfrak{C}) = Q(t - a)\mathbf{F}_{\mathfrak{B},t-a}(\mathfrak{C}) + \mathbf{I}(\mathfrak{C}, t).$$

Here $\mathbf{I}(\mathfrak{C}, t)$ will be called the INERTIAL FORCE acting on \mathfrak{C} due to the change of frame $\{\mathbf{c}, Q, a\}$.

DEFINITION 7: Two dynamical processes $\{\mathfrak{B}, \theta_t, \mathbf{F}_{\mathfrak{B},t}\}$ and $\{\mathfrak{B}, \theta'_t, \mathbf{F}'_{\mathfrak{B},t}\}$ are called EQUIVALENT if there is a change of frame $\{\mathbf{c}, Q, a\}$ such that θ'_t and $\mathbf{F}'_{\mathfrak{B},t}$ are related to θ_t and $\mathbf{F}_{\mathfrak{B},t}$ by (6.3) and (6.4).

The classical analysis of relative motion shows that the inertial force $\mathbf{I}(\mathfrak{C}, t)$ is necessarily of the form

$$(6.5) \quad \mathbf{I}(\mathfrak{C}, t) = \int_{\mathfrak{C}} \mathbf{i}(X, t) dm$$

with

$$(6.6) \quad \mathbf{i}(X, t) = \ddot{\mathbf{c}}(t - a) + 2\dot{V}(t - a)[\mathbf{v}'(X, t) - \dot{\mathbf{c}}(t - a)] \\ + [V^2(t - a) - \dot{V}(t - a)][\theta'_t(X) - \mathbf{c}(t - a)]$$

where \mathbf{v}' is the velocity of the motion θ_t' , and where $V(t)$ is defined by

$$(6.7) \quad V(t) = \dot{Q}(t)Q(t)^{-1}.$$

It is not hard to see that the inertial force \mathbf{I} gives a contribution only to the external body forces and that the contact forces and the mutual forces transform according to (6.2) and hence are objective. The external body forces and the inertial forces cannot be separated from each other in an objective manner. Experience shows that, for the body consisting of the entire solar system, there are frames relative to which the external body forces nearly vanish. These are the classical Galilean frames. Two equivalent dynamical processes really correspond to the same physical process, viewed only from two different frames of reference.

7. Constitutive assumptions. An axiom that characterizes the particular material properties of a body is called a CONSTITUTIVE ASSUMPTION. It restricts the class of dynamical processes the body can undergo. A familiar example is the assumption that the body is rigid. It restricts the possible motions to those in which the distance between any two particles remains unchanged in time. More important for modern continuum mechanics are constitutive assumptions in the form of functional relations between the stress tensor S and the motion θ_t . Such relations are called CONSTITUTIVE EQUATIONS (sometimes also rheological equations of state or stress-strain relations). A classical example is the constitutive equation for linear viscous fluids

$$(7.1) \quad S = (-p + \lambda \operatorname{tr} D)I + 2\mu D,$$

where D is the rate of deformation tensor, I is the unit tensor, p is the pressure, and λ and μ are viscosity constants. A wide variety of constitutive equations have been investigated in recent years⁵, and a general theory of such equations has been developed [2].

Constitutive assumptions are subject to a general restriction:

PRINCIPLE OF OBJECTIVITY: *If a dynamical process is compatible with a constitutive assumption then all processes equivalent to it must also be compatible with this constitutive assumption. In other words, constitutive assumptions must be invariant under changes of frame.*

This principle, although implicitly used by many scientists in the history of mechanics, was stated explicitly first by Oldroyd [3] and was

⁵ A review of the literature and a bibliography is given in [1].

clarified further by the author [4]. It is of great importance in the theory of constitutive equations.

8. Unsolved problems. The axiomatic treatment given here is still too special. It does not cover concentrated forces, contact couples and body couples, sliding, impact, rupture, and other discontinuities, singularities, and degeneracies. It would be desirable to have a universal scheme which covers any conceivable situation.

A more fundamental physical problem is to find a rigorous unified theory of continuum mechanics and thermodynamics. Classical thermodynamics deals only with equilibrium states and hence is not adequate for processes with fast changes of state in time. Such a unified theory should lead to further restrictive conditions on the form of constitutive equations and hence to more definite and realistic theories for special materials. Also, a satisfactory connection with statistical mechanics can be expected only after such a theory has been developed.

References

- [1] NOLL, W., ERICKSEN, J. L. and TRUESDELL, C., *The Non-linear Field Theories of Mechanics*. Article to appear in the Encyclopedia of Physics.
- [2] ———, *A general theory of constitutive equations*. To appear in Archive for Rational Mechanics and Analysis.
- [3] OLDROYD, J. G., *On the formulation of rheological equations of state*. Proceedings of the Royal Society of London (A) 200 (1950), 523–541.
- [4] NOLL, W., *On the Continuity of the Solid and Fluid States*. Journal of Rational Mechanics and Analysis 4 (1955), 3–81.

ZUR AXIOMATISIERUNG DER MECHANIK

HANS HERMES

Universität Münster, Münster in Westfalen, Deutschland

1. Seit Newton ist die Mechanik mehrfach axiomatisiert worden. Es sind Axiome gegeben worden für die Mechanik der Massenpunkte und für die Mechanik der Kontinua, für die nicht-relativistische und für die relativistische Mechanik.

Für die vorliegende Betrachtung sollen diese Unterschiede keine Rolle spielen. Wir wollen uns vielmehr dafür interessieren, welcher Art die Grundbegriffe sind, die in den Axiomensystemen auftreten. Die meisten Axiomatisierungen verwenden u.a. kinematische Grundbegriffe, wie die Begriffe des Ortes, der Geschwindigkeit oder der Beschleunigung. Dabei bezieht man sich entweder auf ein festes System, oder man lässt eine Klasse von Bezugssystemen zu, wobei der Übergang zwischen den einzelnen zugelassenen Bezugssystemen vermittelt wird durch Galilei- bzw. Lorentztransformationen.

Auf die Möglichkeit, kinematische Grundbegriffe zu vermeiden, indem man sie mit Hilfe von Definitionen auf solche zurückführt, die epistemologisch vorangehen, soll hier nicht eingegangen werden.

In den meisten Axiomensystemen (z.B. bei McKinsey, Sugar und Suppes [3]) findet man aber nicht nur rein kinematische Grundbegriffe, sondern es treten in ihnen Grundbegriffe auf, wie die Begriffe der Masse, des Impulses oder der Kraft. Diese Begriffe muss man als typische dynamische oder eigentlich mechanische Begriffe ansehen. Es gibt aber auch Axiomensysteme (z.B. Hermes [2]), welche ausschliesslich mit kinematischen Grundbegriffen auskommen.

Wenn man diese beiden Möglichkeiten ins Auge fasst, wird man sich fragen, welche Gesichtspunkte man anführen kann, die zugunsten der einen oder der anderen Möglichkeit sprechen. Da muss zunächst hervorgehoben werden, dass ein Axiomensystem, welches nicht nur kinematische Grundbegriffe verwendet, viel einfacher ist, als ein Axiomensystem, welches nur kinematische Grundbegriffe enthält. Vom Standpunkt der formalen Eleganz aus werden daher Axiomensysteme stets vorzuziehen sein, die z.B. den Massenbegriff als Grundbegriff enthalten. Ein anderer

Grund, der zugunsten solcher Axiomensysteme spricht, wird am Schluss dieser Nummer genannt.

Gegen die Verwendung des Massenbegriffes und ähnlicher Begriffe als Grundbegriffe in einem Axiomensystem der Mechanik spricht die folgende Ueberlegung: Für einen Physiker sind die Massen, Impulse oder Kräfte nicht unmittelbar gegeben. Er muss diese Grössen vielmehr durch Messungen bestimmen. Eine solche Messung besteht aber letzten Endes in einer Reduktion auf kinematische Begriffe. Wenn man etwa eine Masse mittels einer Federwaage bestimmt, macht man eine Ortsmessung; bestimmt man sie mit Hilfe von Stossgesetzen, so misst man Geschwindigkeiten; oder aber man bedient sich des dritten Newtonschen Axioms und stellt Beschleunigungen fest. Man kann nun den Wunsch haben, der Tatsache, dass ein Physiker auf solche Weise mechanische Grössen mit Hilfe kinematischer Messungen ermittelt, in einer Axiomatisierung dadurch Rechnung zu tragen, dass man den Begriff der Masse und verwandte Begriffe durch Definitionen auf kinematische Begriffe zurückführt, die den physikalischen Messmöglichkeiten entsprechen.

Bei der Bestimmung mechanischer Grössen durch kinematische Messungen muss man die Gültigkeit des einen oder des anderen physikalischen Gesetzes voraussetzen. Bei der Bestimmung der Masse z.B. kann dies das Stossgesetz oder das Gravitationsgesetz sein (vgl. [1]). Eine entsprechende Definition der Masse muss sich der jeweiligen physikalischen Hypothese bedienen. Je nachdem, welche Hypothese man in die Definition der Masse hineinsteckt, kommt man zu verschiedenen und primär unvergleichbaren Theorien der Mechanik. Man sollte dies klar zum Ausdruck bringen und deutlich verschiedene Mechaniken unterscheiden, genau so, wie man sich seit längerem daran gewöhnt hat, von verschiedenen Geometrien zu sprechen.

Jede solche Mechanik ist natürlich eine Idealisierung. Sie ist es insbesondere in folgender Hinsicht: Ein Physiker wird sich bei seinen Messungen natürlich nicht darauf beschränken, z.B. die Masse ausschliesslich mit Hilfe eines einzigen physikalischen Gesetzes zu bestimmen; er wird sich vielmehr vorbehalten, je nach den Umständen das geeignetste Gesetz zu wählen. Eine Axiomatisierung der Mechanik, welche die Masse auf Grund einer einzigen physikalischen Hypothese definiert, bevorzugt dieses Gesetz in einem besonderen Masse. Man wird sich umso eher mit einer solchen Bevorzugung befreunden können, je grundlegender das Gesetz ist, auf welches dabei zurückgegriffen wird.

Die Tatsache, dass man nicht ohne weiteres geneigt sein wird, bei einer

Definition z.B. der Masse ein bestimmtes physikalisches Gesetz zu bevorzugen, mag dazu beigetragen haben, dass viele Autoren es vorziehen, die Masse und andere nicht-kinematische Begriffe bei einer Axiomatisierung der Mechanik als Grundbegriffe zu verwenden.

2. Im folgenden soll berichtet werden über den in [2] unternommenen Versuch, die Mechanik zu axiomatisieren unter Verwendung rein kinematischer Grundbegriffe. Dabei wird zur Definition der Masse zurückgegriffen auf das grundlegende Gesetz der Erhaltung des Impulses bei unelastischen Zusammenstößen. Gleichzeitig sollen hier einige Unvollkommenheiten beseitigt werden, auf welche B. Rosser in seinem Referat [4] hingewiesen hat (vgl. auch die Korrekturen im *Anhang*). In der genannten Abhandlung [2] ist die relativistische Kontinuumsmechanik aufgebaut worden. Da es im folgenden nur auf die Grundgedanken ankommt, soll hier die nicht-relativistische Punktmechanik axiomatisiert werden, was im Einzelnen wesentlich einfacher ist.

Zunächst einige Vorbemerkungen zur Symbolisierung. Bei der Axiomatisierung wird eine Theorie der reellen Zahlen vorausgesetzt. Die eigentlichen mechanischen Aussagen werden in einer Stufenlogik wiedergegeben, wobei auf der untersten Stufe zwei Sorten von Individuenvariablen verwendet werden. Die Individuenvariablen τ_1, τ_2, \dots beziehen sich auf reelle Zahlen, die Individuenvariablen x, y, \dots auf momentane Massenpunkte. Ein *momentaner Massenpunkt* ist ein zu einem bestimmten Zeitpunkt betrachteter Massenpunkt, also ein zeitlicher Schnitt durch die Weltlinie eines Massenpunktes. Man kann auf die explizite Verwendung des Begriffs eines Massenpunktes verzichten, wenn man einen Massenpunkt auffasst als eine grösste Klasse „zusammengehöriger“ momentaner Massenpunkte. Die Zusammengehörigkeit momentaner Massenpunkte bedeutet ihre Zugehörigkeit zu einem und demselben Massenpunkt. Zusammengehörige momentane Massenpunkte sollen nach Levin *genidentisch* genannt werden.

Orte und Zeiten momentaner Massenpunkte werden durch Bezugssysteme festgelegt. Ein *Bezugssystem* Σ , wie es in der Mechanik verwendet wird, kann aufgefasst werden als eine fünfstellige Relation zwischen reellen Zahlen $\tau_1, \tau_2, \tau_3, \tau_4$ und momentanen Massenpunkten x . $\Sigma\tau_1\tau_2\tau_3\tau_4x$ besagt, dass x im Bezugssystem Σ die Raumkoordinaten τ_1, τ_2, τ_3 und die Zeitkoordinate τ_4 besitzt. Häufig wird für $\Sigma\tau_1\tau_2\tau_3\tau_4x$ die abkürzende Bezeichnung $\Sigma\tau x$ verwendet.

Das Axiomensystem enthält zwei Grundbegriffe, nämlich die zwei-

stellige Relation G und die Klasse \mathfrak{B} . Gxy soll bedeuten, dass die momentanen Massenpunkte x und y genidentisch sind, d.h., dass sie zu demselben Massenpunkt gehören. $\mathfrak{B}\Sigma$ besage, dass Σ ein *galileisches Bezugssystem* (*Inertialsystem*) ist.

3. Nach diesen Vorbereitungen sollen die Axiome formuliert werden. Der Einfachheit halber werden die Axiome angegeben unter Verwendung der Konvention, dass frei vorkommende Variablen $\tau_1, \tau_2, \dots, x, y, \dots, \Sigma, \dots$ generalisiert gedacht werden.

Das erste Axiom sagt aus, dass G eine Äquivalenzrelation ist:

AXIOM 1.1. Gxx

AXIOM 1.2. $Gxy \rightarrow Gyx$

AXIOM 1.3. $Gxy \wedge Gyz \rightarrow Gxz$

Axiom 2.1 bringt zum Ausdruck, dass die Koordinaten eines momentanen Massenpunktes x in jedem Bezugssystem Σ eindeutig festgelegt sind. Axiom 2.2 besagt, dass genidentische momentane Massenpunkte x, y identisch sind, wenn sie in einem Bezugssystem Σ dieselbe Zeitkoordinate besitzen. In Axiom 2.3 wird gefordert, dass ein Massenpunkt eine „unendliche Lebensdauer“ besitzt.

AXIOM 2.1. $\mathfrak{B}\Sigma \wedge \Sigma \tau_1 \tau_1 x \wedge \Sigma \tau_2 \tau_2 x \rightarrow \tau_1 = \tau_2 \wedge \tau_1 = \tau_2$

AXIOM 2.2. $\mathfrak{B}\Sigma \wedge Gx_1 x_2 \wedge \Sigma \tau_1 \tau_1 x_1 \wedge \Sigma \tau_2 \tau_2 x_2 \rightarrow x_1 = x_2$

AXIOM 2.3. $\mathfrak{B}\Sigma \rightarrow \forall \tau \forall y (Gxy \wedge \Sigma \tau y)$

Der Zusammenhang zwischen den verschiedenen Koordinatensystemen wird hergestellt mittels der sog. *Galileitransformationen*. *gal* Γ bedeute, dass Γ eine Galileitransformation ist. Der Begriff der Galileitransformation ist bekannt. Es handelt sich bei jeder solchen Transformation um einen speziellen Automorphismus des gesamten reellen vierdimensionalen Raumes. Wenn man die Koordinaten aller momentanen materiellen Punkte in einem Inertialsystem Σ einer derartigen Galileitransformation Γ unterwirft, so erhält man neue Koordinaten. Dass diese Zuordnung wieder ein Inertialsystem ist, wird in Axiom 3.1 gefordert. Dass je zwei Inertialsysteme in dieser Weise zusammenhängen, wird in Axiom 3.2 ausgesagt. Hierbei muss man beachten (vgl. hierzu die Rossersche Kritik an der ursprünglichen Formulierung des entsprechenden Axioms A.4.5

in [2]), dass nicht angenommen wird, dass der gesamte dreidimensionale Raum mit Massenpunkten besetzt ist. Zwei Inertialsysteme Σ_1 und Σ_2 liefern daher Transformationen nur für solche Quadrupel reeller Zahlen, zu denen es momentane Massenpunkte gibt, welche in Σ_1 diese Quadrupel als Koordinaten haben. Man darf daher nur verlangen, dass die auf diese Weise gewonnene Koordinatentransformation in einer Galileitransformation enthalten ist.

In den beiden folgenden Axiomen treten zwei „Verkettungen“ auf, welche zunächst erklärt werden sollen:

DEFINITION $(\Gamma/\dot{\Sigma})\tau\tau x$ bedeute $\bigvee_{\tau'} \bigvee_{\tau'} (\Gamma\tau\tau'\tau' \wedge \Sigma\tau'\tau'x)$

DEFINITION $(\Sigma_1/\dot{\Sigma}_2)\tau\tau\tau' x$ bedeute $\bigvee_x (\Sigma_1\tau\tau x \wedge \Sigma_2\tau'\tau'x)$

Damit formulieren wir

AXIOM 3.1. $\mathfrak{B}\Sigma \wedge gal\Gamma \rightarrow \mathfrak{B}(\Gamma/\dot{\Sigma})$

AXIOM 3.2. $\mathfrak{B}\Sigma_1 \wedge \mathfrak{B}\Sigma_2 \rightarrow \bigvee_{\Gamma} (gal\Gamma \wedge \Sigma_1/\dot{\Sigma}_2 \subset \Gamma)$

Wir wollen die Massen aus Geschwindigkeiten bei Stossversuchen ablesen. Legen wir ein Inertialsystem Σ zugrunde, so ist die Geschwindigkeit eines momentanen Massenpunktes x_0 gegeben als $\lim_{\tau \rightarrow \tau_0} (\tau - \tau_0)/(\tau - \tau_0)$; dabei sei $\Sigma\tau_0\tau_0x_0$ und es gelte $\Sigma\tau\tau x$ für denjenigen momentanen Massenpunkt x , der mit x_0 genidentisch ist, und dem in Σ die Zeitkoordinate τ zukommt. Die Existenz und Eindeutigkeit von x ergibt sich aus den Axiomen 2.3 und 2.2. Wir wollen im folgenden unelastische Stösse betrachten. Es ist am einfachsten anzunehmen, dass solche Stösse momentan erfolgen und dass sich dabei die Geschwindigkeiten der beteiligten Massenpunkte unstetig ändern. Wir werden daher von Geschwindigkeiten unmittelbar vor dem Stosse oder kurz *Vorgeschwindigkeiten* und entsprechend von *Nachgeschwindigkeiten* reden. $Vel_- \Sigma\mathfrak{v}\tau x$ soll besagen, dass im Inertialsystem Σ der momentane Massenpunkt x zur Zeit τ die Vorgeschwindigkeit \mathfrak{v} besitzt. Analog sei $Vel_+ \Sigma\mathfrak{v}\tau x$ eingeführt. Wir wollen hier der Kürze halber auf die explizite Definition von Vel_- und Vel_+ verzichten und das nächste Axiom nur umgangssprachlich formulieren:

AXIOM 4. *Die Massenpunkte besitzen eine stückweise stetige Geschwindigkeit; an den Sprungstellen existieren wenigstens die Grenzwerte von links bzw. von rechts (Vor- bzw. Nachgeschwindigkeit).*

4. Bei einem Stoss treffen sich zwei Massenpunkte an derselben Stelle. Dies soll hier (im Gegensatz zu der zitierten Abhandlung) ausdrücklich zugelassen werden, um die Stossgesetze so einfach wie möglich darstellten zu können. Es muss jedoch bemerkt werden, dass damit das Prinzip der Undurchdringlichkeit der Materie geopfert wird. (In der zitierten Abhandlung [2] wird keine derartige Annahme gemacht.)

Ein unelastischer Stoss ist dadurch gekennzeichnet, dass die beiden beteiligten Massenpunkte unmittelbar nach dem Stoss dieselbe Geschwindigkeit haben. Diese Geschwindigkeit kann bei Wahl eines geeigneten Bezugssystems Σ als 0 angenommen werden. Es werde gefordert, dass die Geschwindigkeiten unmittelbar vor dem Stoss von 0 verschieden sind. Schliesslich muss noch verlangt werden, dass nur die beiden betrachteten Massenpunkte am Stoss beteiligt sind, d.h., dass sich zur Zeit des Stosses kein dritter Massenpunkt am Stossort befindet. *Stoss* $\Sigma \tau x_1 x_2$ soll bedeuten, dass die zu den momentanen Massenpunkten x_1, x_2 gehörenden Massenpunkte zu der in Σ gemessenen Zeit τ einen Stoss erleiden, bei welchem sie (in Σ gemessen) zur Ruhe kommen.

DEFINITION: *Stoss* $\Sigma \tau x_1 x_2 =_{\text{Df}} \exists \Sigma \wedge \bigvee_{\tau} (\Sigma \tau x_1 \wedge \Sigma \tau x_2)$

$$\wedge Vel_{+} \Sigma v \tau x_1 \wedge Vel_{+} \Sigma v \tau x_2$$

$$\wedge \bigvee_{v_1 v_2} (Vel_{-} \Sigma v_1 \tau x_1 \wedge Vel_{-} \Sigma v_2 \tau x_2 \wedge v_1 \neq 0 \wedge v_2 \neq 0)$$

$$\wedge \bigwedge_{y \tau} (\Sigma \tau y \wedge \Sigma \tau x_1 \rightarrow y = x_1 \vee y = x_2)$$

Für den unelastischen Stoss gilt das Gesetz der Impulserhaltung. Sind m_1 bzw. m_2 die Massen der beteiligten Massenpunkte und v_1 bzw. v_2 die in Σ gemessenen Geschwindigkeiten vor dem Stoss, so hat man $m_1 v_1 + m_2 v_2 = 0$. Damit ergibt sich die Möglichkeit, dass Massenverhältnis α aus dem Verhältnis der Geschwindigkeitsbeträge zu ermitteln. *Masse* $\alpha x x_0$ soll inhaltlich besagen, dass die Masse von x α -mal so gross ist wie die Masse des Vergleichsmassenpunktes x_0 .

Sind x und x_0 genidentisch, so wird ein Stossversuch illusorisch. In diesem Fall soll das Massenverhältnis per definitionem gleich eins gesetzt werden. Wir kommen damit zu der grundlegenden

DEFINITION: *Masse* $\alpha x x_0 =_{\text{Df}} \bigvee_{\Sigma \tau} \bigvee_{v} \bigvee_{v_0} \bigvee_{v} \bigvee_{v_0} (Gx y \wedge Gx_0 y_0 \wedge \text{Stoss } \Sigma \tau y y_0$

$$\wedge Vel_{-} \Sigma v \tau y \wedge Vel_{-} \Sigma v_0 \tau y_0 \wedge \alpha \cdot |v| = |v_0|) \vee (Gx x_0 \wedge \alpha = 1)$$

Im folgenden sollen weitere Axiome formuliert werden, welche sich auf die Existenz und die Eindeutigkeit des Massenverhältnisses beziehen. Da es hier nur auf eine prinzipielle Diskussion ankommt, soll kein Wert darauf gelegt werden, die Axiome so schwach wie möglich zu formulieren. Für einen vollständigen Aufbau der Mechanik ist es natürlich erforderlich, über die genannten Axiome hinaus noch weitere zu fordern, welche sich z.B. auf die Gültigkeit des Impulssatzes beziehen. Ausserdem müsste u.a. der Begriff der Kraft eingeführt werden. Hierzu soll auf [2] verwiesen werden.

Zunächst formulieren wir ein Axiom, welches zum Ausdruck bringt, dass es sich bei der soeben eingeführten Masse um ein wirkliches Verhältnis handelt.

AXIOM 5. $Masse\ \alpha yz \wedge Masse\ \beta zx \wedge Masse\ \gamma xy \rightarrow \alpha\beta\gamma = 1$

Wir formulieren nun einige einfache Sätze. Satz 5 sagt aus, dass das Massenverhältnis eindeutig ist, d.h. nur von den beteiligten Massenpunkten abhängt.

SATZ 1: $Stoss\ \Sigma\tau x_1x_2 \rightarrow Stoss\ \Sigma\tau x_2x_1$

SATZ 2: $Masse\ \alpha x x_0 \rightarrow \alpha \neq 0$

SATZ 3: $Masse\ \alpha x x_0 \rightarrow Masse\ \frac{1}{\alpha} x_0 x$

SATZ 4: $Masse\ \alpha x x_0 \wedge Gxy \wedge Gx_0y_0 \rightarrow Masse\ \alpha y y_0$

SATZ 5: $Masse\ \alpha x x_0 \wedge Masse\ \beta y y_0 \wedge Gxy \wedge Gy_0y_0 \rightarrow \alpha = \beta$

BEWEIS: Satz 1 folgt unmittelbar aus der Definition des Stosses. Satz 2 ergibt sich daraus, dass nach der Definition des Stosses die in Frage kommenden Vorgeschwindigkeiten von 0 verschieden sind. Satz 3 ergibt sich aus Satz 1 und Satz 2. Satz 4 folgt aus der Massendefinition. Satz 5 zeigt man so: Zunächst hat man $Masse\ \alpha y y_0$ nach Satz 4 und damit $Masse\ \frac{1}{\alpha} y_0 y$ nach Satz 3. Wegen Gyy gilt $Masse\ 1yy$. Nun hat man $Masse\ \beta y y_0 \wedge Masse\ \frac{1}{\alpha} y_0 y \wedge Masse\ 1yy$, also $\beta \cdot \frac{1}{\alpha} \cdot 1 = 1$ nach Axiom 5.

Das letzte Axiom, welches hier diskutiert werden soll, betrifft die

Existenz des Massenverhältnisses:

AXIOM 6. $\forall \text{ Masse } \alpha \exists x_0$
 α

Dieses Axiom bringt zum Ausdruck, dass je zwei verschiedene Massenpunkte mindestens einmal unelastisch zusammenstossen. Das ist eine sehr starke Forderung. (Man könnte diese Forderung abschwächen, indem man nur verlangte, dass es zu je zwei verschiedenen Massenpunkten eine endliche Kette von Massenpunkten gibt, von denen je zwei aufeinanderfolgende Massenpunkte irgendwann unelastisch zusammenstossen, und wenn man die Definition des Massenverhältnisses entsprechend modifizierte. Aber auch ein derartig modifiziertes Axiom würde eine starke Forderung aussprechen.) Zu diesem Axiom (bzw. zu dem analogen Axiom 8.1 in [2]) sagt Rosser in [4], dass verlangt wird, dass die Massenpunkte "behave in certain very peculiar fashions". Axiom 6 mag jedoch weniger befremdlich erscheinen, wenn man sich vergegenwärtigt, dass es entstanden ist als eine Formulierung der idealisierten Vorstellung, dass Physiker Massen durch Stossversuche bestimmen können.

Zugunsten der angegebenen Formulierung mag auch noch ein analoger Sachverhalt aus der Geometrie herangezogen werden. Ein geometrisches Axiom sagt aus, dass es zu je zwei voneinander verschiedenen Punkten eine Gerade gibt, welche diese beiden Punkte verbindet. Die geometrischen Axiome geben wie die Axiome der Mechanik ursprünglich physikalische Sachverhalte wieder. Denkt man an eine Realisierung der Geraden etwa durch gespannte Seile, so soll durch das genannte Axiom zum Ausdruck gebracht werden, dass es zwischen je zwei Raumpunkten eine derartige Seilverbindung gibt. Dies ist völlig analog zu der Forderung, dass je zwei Massenpunkte im Laufe der Zeit einen unelastischen Zusammenstoss erleiden.

Das soeben betrachtete geometrische Beispiel gibt aber auch einen Hinweis darauf, wie man zu einer plausiblen Abschwächung der betrachteten Axiome kommen kann. Man könnte nämlich sagen, dass die Möglichkeit besteht, zwei beliebige Raumpunkte durch ein gespanntes Seil zu verbinden. Damit wäre das Axiom streng genommen nur eine Möglickeitsaussage. Entsprechend könnte man das mechanische Axiom 6 so abschwächen, dass man nur verlangt, dass es möglich ist, dass je zwei Massenpunkte irgendwann unelastisch zusammenstossen.

Eine andere Methode, einer so starken Formulierung, wie sie Axiom 6 darstellt, zu entgehen, besteht darin, den Massenbegriff durch eine Reduktion (nach Carnap) auf kinematische Begriffe zurückzuführen.

ANHANG: *Corrigenda zu [2]*:

S. 10, Z. 12 v.u. lies: „((0000)*)”. S. 10, Z. 5. v.u. lies: „ $f(x) \rightarrow g(x)$ ”.
 S. 13 füge ein: „A4.3': $\Sigma \dot{\Sigma} \in Bzs$ ”. S. 13, A4.5 lies: „ $\dot{I}(I \in lo \wedge \Sigma_1 / \Sigma_2 \supset I)$ ”.
 S. 28, Z. 8 v.o. statt: „vergleichbaren” lies: „genidentischen”. S. 31, Z. 6 v.o. statt „*Isol*” lies „*Is*”.

Bibliographie

- [1] ADAMS, E. W., *The foundations of rigid body mechanics and the derivation of its law from those of particle mechanics*. This volume, pp. 250–265.
- [2] HERMES, H., *Eine Axiomatisierung der allgemeinen Mechanik*. Forschungen zur Logik und zur Grundlegung der exakten Wissenschaften. Neue Folge, Heft 3, Leipzig 1938, 48 S.
- [3] MCKINSEY, J. C. C., SUGAR, A. C. and SUPPES, P., *Axiomatic Foundations of Classical Particle Mechanics*. Journal of Rational Mechanis and Analysis, vol. 2 (1953), pp. 253–272.
- [4] ROSSER, B., *Review of HERMES [1]*. Journal of Symbolic Logic, vol. 3 (1938), pp. 119–120.

AXIOMS FOR RELATIVISTIC KINEMATICS WITH OR WITHOUT PARITY

PATRICK SUPPES

Stanford University, Stanford, California, U.S.A.

1. Introduction. The primary aim of this paper is to give an elementary derivation of the Lorentz transformations, without any assumptions of continuity or linearity, from a single axiom concerning invariance of the relativistic distance between any two space-time points connected by an inertial path. The concluding section considers extensions of the theory of relativistic kinematics which will destroy conservation of temporal parity, that is, extensions which are not invariant under time reversals.

It is philosophically and empirically interesting that the Lorentz transformations can be derived without any extraneous assumptions of continuity or differentiability. In a word, the single assumption needed for relativistic kinematics is that all observers at rest in inertial frames get identical measurements of relativistic distances along inertial paths when their measuring instruments have identical calibrations. Note that it is a consequence and *not* an assumption that these observers are moving with a uniform velocity with respect to each other. Granted the possibility of perfect measurements everywhere of relativistic intervals, this single axiom isolates in a precise way the narrow operational basis needed for the special theory of relativity.

Prior to any search of the literature it would seem that this result would be well-known, but I have not succeeded in finding the proof anywhere. Every physics textbook on relativity makes a linearity assumption at the minimum. In geometrical discussions of indefinite quadratic forms it is often remarked that the relativistic interval is invariant under the Lorentz group, but it is not proved that it is invariant under no wider group, which is the main fact established here. Some further remarks in this connection are made at the end of Section 2.

2. Primitive Notions and Single Axiom. Our single initial axiom for relativistic kinematics is based on three primitive notions, each of which has a simple physical interpretation. The first notion is an arbitrary set X

interpreted as the set of *physical space-time points*. The second notion is a non-empty family \mathfrak{F} of one-one functions mapping X onto R_4 , the set of all ordered quadruples of real numbers. (Thus X must have the power of the continuum.) Intuitively each function in \mathfrak{F} represents an *inertial space-time frame of reference*, or, more explicitly, a space-time measuring apparatus at rest in an inertial frame. If $x \in X$, $f \in \mathfrak{F}$, and $f(x) = \langle x_1, x_2, x_3, t \rangle$ then x_1 , x_2 , and x_3 are the three orthogonal spatial coordinates of the point x , and t the time coordinate, with respect to the frame f . For a more explicit formal notation, $f_i(x)$ is the i th coordinate of the space-time point x with respect to the frame f , for $i = 1, \dots, 4$. The third primitive notion is a positive number c , which is to be interpreted as the *speed of light*.

It is convenient to have a notation for the *relativistic distance* with respect to a frame f between any two space-time points x and y .

DEFINITION 1. If $x, y \in X$ and $f \in \mathfrak{F}$ then

$$I_f(xy) = \sqrt{\sum_{i=1}^3 [f_i(x) - f_i(y)]^2 - c^2[f_4(x) - f_4(y)]^2}.$$

(We always take the square-root with positive sign.) If f is an inertial frame, then (i) $I_f(xy) = 0$ if x and y are connected by a light line, (ii) $I_f^2(xy) < 0$ if x and y lie on an inertial path (the square is negative since $I_f(xy)$ is imaginary); (iii) $I(xy) > 0$ if x and y are separated by a "space-like" interval. We use (ii) for a formal definition.

DEFINITION 2. If $x, y \in X$ and $f \in \mathfrak{F}$ then x AND y LIE ON AN INERTIAL PATH WITH RESPECT TO f if and only if $I_f^2(xy) < 0$.

It will also occasionally be useful to characterize inertial paths in terms of their speed. We may do this informally as follows. By the *slope* of a line α in R_4 , whose projection on the 4th coordinate (the time coordinate) is a non-degenerate segment, we mean the three-dimensional vector W such that for any two distinct points $\langle Z_1, t_1 \rangle$ and $\langle Z_2, t_2 \rangle$ of α

$$W = \frac{Z_1 - Z_2}{t_1 - t_2}.$$

By the *speed* of α we mean the non-negative number $|W|$. An *inertial path*

is a line in R_4 whose speed is less than c ; and a *light line* is of course a line whose speed is c .

The single axiom we require is embodied in the following definition.

DEFINITION 3. *A system $\mathfrak{K} = \langle X, \mathfrak{F}, c \rangle$ is a COLLECTION OF RELATIVISTIC FRAMES if and only if for every x, y in X , whenever x and y lie on an inertial path with respect to some frame in \mathfrak{F} , then for all f, f' in \mathfrak{F}*

$$(1) \quad I_f(xy) = I_{f'}(xy).$$

I originally formulated this invariance axiom so as to require that equation (1) hold for *all* space-time points x and y , that is, without restricting them to lie on an inertial path (with respect to some frame in \mathfrak{F}). Walter Noll pointed out to me that with this stronger axiom no physically motivated arguments of the kind given below are required to prove that any two frames in \mathfrak{F} are related by a linear transformation; a relatively simple algebraic argument may be given to show this.

On the other hand, when the invariance assumption is restricted, as it is here, to distances between points on inertial paths, the line of argument formalized in the theorems of the next section seems necessary. This restriction to pairs of points on inertial paths is physically natural because their distances $I_f(xy)$ are more susceptible to direct measurements than are the distances of points separated by a space-like interval (i.e., $I_f(xy) > 0$).

3. Theorems. In proving the main result that any two frames in \mathfrak{F} are related by a Lorentz transformation, some preliminary definitions, theorems and lemmas will be useful. We shall use freely the geometrical language appropriate to Euclidean four-dimensional space with the ordinary positive definite quadratic form.

THEOREM 1. *If $k \geq 0$ and $f(x) - f(y) = k[f(u) - f(v)]$ then $I_f(xy) = kI_f(uv)$.*

PROOF: If $k = 0$, the theorem is immediate. So we need to consider the case for which $k > 0$. It follows from the hypothesis of the theorem that

$$(1) \quad x_i - y_i = k(u_i - v_i) \text{ for } i = 1, \dots, 4,$$

where, for brevity here and subsequently, when we are considering a fixed element f of \mathfrak{F} , $f_i(x) = x_i$, etc. Using (1) and Definition 1 we then

have:

$$\begin{aligned}
 I_f(xy) &= \sqrt{\sum_{i=1}^3 (x_i - y_i)^2 - c^2(x_4 - y_4)^2} \\
 &= \sqrt{\sum_{i=1}^3 k^2(u_i - v_i)^2 - c^2k^2(u_4 - v_4)^2} \\
 &= kI_f(uv). \qquad \text{Q.E.D.}
 \end{aligned}$$

In the next theorem we use the notion of *betweenness* in a way which is meant not to exclude identity with one of the end points.

THEOREM 2. *If the points $f(x)$, $f(y)$ and $f(z)$ are collinear and $f(y)$ is between $f(x)$ and $f(z)$ then*

$$I_f(xy) + I_f(yz) = I_f(xz).$$

PROOF: Extending our subscript notation, let $f(x) = \mathbf{x}$, etc. Since the three points \mathbf{x} , \mathbf{y} and \mathbf{z} are collinear, and \mathbf{y} is between \mathbf{x} and \mathbf{z} , there is a number k such that $0 \leq k \leq 1$ and

$$(1) \qquad \mathbf{y} = k\mathbf{x} + (1 - k)\mathbf{z},$$

whence

$$\mathbf{y} - \mathbf{z} = k(\mathbf{x} - \mathbf{z}),$$

and thus by Theorem 1

$$(2) \qquad I_f(yz) = kI_f(xz).$$

By adding and subtracting \mathbf{x} from the right-hand side of (1), we get:

$$\mathbf{y} = k\mathbf{x} + (1 - k)\mathbf{z} + \mathbf{x} - \mathbf{x},$$

whence

$$\mathbf{x} - \mathbf{y} = (1 - k)(\mathbf{x} - \mathbf{z}),$$

and thus by virtue of Theorem 1 again,

$$(3) \qquad I_f(xy) = (1 - k)I_f(xz).$$

Adding (2) and (3) we obtain the desired result:

$$I_f(xy) + I_f(yz) = I_f(xz). \qquad \text{Q.E.D.}$$

Our next objective is to prove a partial converse of Theorem 2. Since the notion of Lorentz transformation is needed in the proof, we introduce

the appropriate formal definitions at this point. \mathcal{I} is the identity matrix of the necessary order.

DEFINITION 4. *A matrix \mathcal{A} (of order 4) is a LORENTZ MATRIX if and only if there exist real numbers β , δ , a three-dimensional vector U , and an orthogonal matrix \mathcal{E} of order 3 such that*

$$\begin{aligned}\beta^2 \left(1 - \frac{U^2}{c^2}\right) &= 1 \\ \delta^2 &= 1 \\ \mathcal{A} &= \begin{pmatrix} \mathcal{E} & 0 \\ 0 & \delta \end{pmatrix} \begin{pmatrix} \mathcal{I} + \frac{\beta - 1}{U^2} U^* U & -\frac{\beta U^*}{c^2} \\ -\beta U & \beta \end{pmatrix}.\end{aligned}$$

(In this definition and elsewhere, if A is a matrix, A^* is its transpose, and vectors like U are one-rowed matrices — thus U^* is a one-column matrix.) The physical interpretation of the various quantities in Definition 1 should be obvious. The number β is the *Lorentz contraction factor*. When $\delta = -1$, we have a reversal of the direction of time. The matrix \mathcal{E} represents a *rotation* of the spatial coordinates, or a rotation followed by a reflection. The vector U is the *relative velocity* of the two frames of reference. For future reference it may be noted that every Lorentz matrix is non-singular.

DEFINITION 5. *A Lorentz transformation is a one-one function φ mapping R_4 onto itself such that there is a Lorentz matrix \mathcal{A} and a 4-dimensional vector B so that for all Z in R_4*

$$\varphi(Z) = Z\mathcal{A} + B.$$

The physical interpretation of the vector B is clear. Its first three coordinates represent a translation of the origin of the spatial coordinates, and its last coordinate a translation of the time origin. Definition 5 makes it clear that every Lorentz transformation is a nonsingular affine transformation of R_4 , a fact which we shall use in several contexts. The important consideration for the proof of Theorem 3 is that affine transformations preserve the collinearity of points.

THEOREM 3. *If any two of the three points x , y , z are distinct and lie on an inertial path with respect to f and if $I_f(xy) + I_f(yz) = I_f(xz)$, then the points $f(x)$, $f(y)$ and $f(z)$ are collinear, and $f(y)$ is between $f(x)$ and $f(z)$.*

PROOF: Three cases naturally arise.

Case 1. $I^2(xy) < 0$. In this case the line segment $f(x) - f(y)$ is an inertial path segment from x to y , and there exists a Lorentz transformation φ which will transform the segment $f(x) - f(y)$ to "rest", that is, more precisely, φ may be chosen so as to transform f to a frame f' , which need not be a member of \mathfrak{F} , such that the spatial coordinates of x and y are at the origin, the time coordinate of x is zero, and z has but one spatial coordinate, by appropriate spatial rotation. That is, we have:

$$\begin{aligned}f'(x) &= \langle 0, 0, 0, 0 \rangle, \\f'(y) &= \langle 0, 0, 0, y'_4 \rangle, \\f'(z) &= \langle z'_1, 0, 0, z'_4 \rangle.\end{aligned}$$

We shall prove that $f'(x)$, $f'(y)$ and $f'(z)$ are collinear. Since φ is non-singular and affine, its inverse φ^{-1} exists and is affine, whence collinearity is preserved in transforming from f' back to f .

It is a familiar fact that the relativistic intervals $I_f(xy)$, $I_f(yz)$ and $I_f(xz)$ are Lorentz invariant and thus have the same value with respect to f' as f . Consequently, from the additive hypothesis of the theorem, we have:

$$(1) \quad \sqrt{-c^2 y_4'^2} + \sqrt{z_1'^2 - c^2(y_4' - z_4')^2} = \sqrt{z_1'^2 - c^2 z_4'^2}.$$

Squaring both sides of (1), then cancelling and rearranging terms, we obtain:

$$(2) \quad \sqrt{-y_4'^2} \cdot \sqrt{z_1'^2 - c^2(y_4' - z_4')^2} = c y_4'(y_4' - z_4').$$

If $y_4' = 0$, then x and y are identical, contrary to the hypothesis that $I^2(xy) < 0$. Taking then $y_4' \neq 0$, dividing it out in (2), squaring both sides and cancelling, we infer:

$$-z_1'^2 = 0,$$

whence

$$z_1' = 0,$$

which establishes the collinearity in f' of the three points, since their spatial coordinates coincide, and obviously $f'(y)$ is between $f'(x)$ and $f'(z)$.

Case 2. $I_f^2(yz) < 0$. Proof similar to Case 1.

Case 3. $I_f^2(xz) < 0$. By an argument similar to that given for Case 1, we may go from f to a frame f' by a Lorentz transformation which will

transform the inertial segment $f(x) - f(z)$ to "rest." That is, we obtain:

$$f'(x) = \langle 0, 0, 0, 0 \rangle,$$

$$f'(y) = \langle y'_1, 0, 0, y'_4 \rangle,$$

$$f'(z) = \langle 0, 0, 0, z'_4 \rangle.$$

Then by the additive hypothesis of the theorem:

$$(3) \quad \sqrt{y_1'^2 - c^2 y_4'^2} + \sqrt{y_1'^2 - c^2 (y_4' - z_4')^2} = \sqrt{-c^2 z_4'^2}.$$

Proceeding as before, by squaring and cancelling, we obtain from (3):

$$(4) \quad \sqrt{-c^2 z_4'^2} \cdot \sqrt{y_1'^2 - c^2 y_4'^2} = -c^2 y_4' z_4'.$$

Squaring again and cancelling yields:

$$(5) \quad y_1'^2 z_4'^2 = 0.$$

There are now two possibilities to consider: either $y_1' = 0$ or $z_4' = 0$. If the former is the case, then the three points are collinear in R_4 , for they are all three placed at the origin of the spatial coordinates. On the other hand, if $z_4' = 0$, then x and z are identical points, contrary to hypothesis. Again it is obvious that $f'(y)$ is between $f'(x)$ and $f'(z)$. Q.E.D.

That a full converse of Theorem 2 cannot be proved, in other words that the additive hypothesis

$$I_f(xy) + I_f(yz) = I_f(xz)$$

does not imply collinearity, is shown by the following counterexample:

$$f(x) = \langle 0, 0, 0, 0 \rangle,$$

$$f(y) = \langle 1, 1, 0, 0 \rangle$$

$$f(z) = \langle \sqrt{2c}, 0, 0, 1 \rangle.$$

Clearly, $f(x)$, $f(y)$ and $f(z)$ are not collinear in R_4 , but $I_f(xy) + I_f(yz) = I_f(xz)$, that is,

$$(1) \quad \sqrt{2} + \sqrt{(1 - \sqrt{2}c)^2 + 1 - c^2} = \sqrt{2c^2 - c^2}.$$

For, simplifying and rearranging (1), we see it is equivalent to:

$$(2) \quad \sqrt{2 - 2\sqrt{2}c + c^2} = c - \sqrt{2}$$

and the left-hand of (2) is simply

$$\sqrt{(c - \sqrt{2})^2} = c - \sqrt{2}.$$

(It may be mentioned that the full converse of Theorem 2 does hold for R_2 , that is, when there is a restriction to one spatial dimension.)

We now want to prove some theorems about properties which are invariant in \mathfrak{F} . Formally, a property is *invariant* in \mathfrak{F} if and only if it holds or does not hold uniformly for every member f of \mathfrak{F} . Thus to say that the property of a line being an inertial path is invariant in \mathfrak{F} means that a line with respect to f in \mathfrak{F} , is an inertial path with respect to f if and only if it is an inertial path with respect to every f' in \mathfrak{F} . All geometric objects referred to here are with respect to the frames in \mathfrak{F} .

THEOREM 4. *The property of being the midpoint of a finite segment of an inertial path is invariant in \mathfrak{F} .*

PROOF: Suppose x , y and z lie on an inertial path with respect to f and

$$(1) \quad f(y) = \frac{1}{2}f(x) + \frac{1}{2}f(z),$$

and thus

$$f(y) - f(x) = \frac{1}{2}[f(z) - f(x)].$$

Consequently by virtue of Theorem 1

$$(2) \quad I_f(xy) = \frac{1}{2}I_f(xz)$$

and similarly

$$(3) \quad I_f(yz) = \frac{1}{2}I_f(xz),$$

whence

$$(4) \quad I_f(xy) + I_f(yz) = I_f(xz).$$

Now by the invariance axiom of Definition 3, for any f' in \mathfrak{F}

$$I_{f'}(xy) = I_f(xy)$$

$$I_{f'}(yz) = I_f(yz)$$

$$I_{f'}(xz) = I_f(xz).$$

Substituting these identities in (4) we obtain:

$$I_{f'}(xy) + I_{f'}(yz) = I_{f'}(xz).$$

Thus by virtue of Theorem 3, $f'(x)$, $f'(y)$ and $f'(z)$ are collinear with $f'(y)$ between $f'(x)$ and $f'(z)$. Moreover, since by the invariance axiom (2) and (3) hold for f' , we conclude $f'(y)$ is actually the midpoint. Q.E.D.

This proof is easily extended to show that the property of being an inertial path is invariant in \mathfrak{F} , but we do not directly need this fact. We next want to show that this midpoint property is invariant for arbitrary segments. In view of the counterexample following Theorem 3 it is evident that a direct proof in terms of the relativistic intervals cannot be given. The method we shall use consists essentially of constructing a parallelogram whose sides are segments of inertial paths. A similar but somewhat more complicated proof is given in Rubin and Suppes [3].

THEOREM 5. *The property of being the midpoint of an arbitrary finite segment is invariant in \mathfrak{F} .*

PROOF: Let $A = \langle Z_1, t_1 \rangle$ and $B = \langle Z_2, t_2 \rangle$ where A is an arbitrary segment in R_4 . (The points A to G defined here are with respect to f in \mathfrak{F} .) For definiteness assume $t_1 \geq t_2$. We set

$$Z_0 = \frac{Z_1 + Z_2}{2}$$

and we choose t_0 and t_3 so that

$$t_0 < t_2 - \frac{|Z_1 - Z_2|}{2c},$$

$$t_3 > t_1 + \frac{|Z_1 - Z_2|}{2c},$$

$$|A - \langle Z_0, t_3 \rangle| = |\langle Z_0, t_0 \rangle - B|,$$

$$|A - \langle Z_0, t_0 \rangle| = |\langle Z_0, t_3 \rangle - B|.$$

We now let (see Figure 1)

$$C = \langle Z_0, t_0 \rangle, \quad D = \langle Z_0, t_3 \rangle,$$

$$E = \frac{A + B}{2}, \quad F = \frac{B + D}{2},$$

$$G = \frac{A + C}{2}.$$

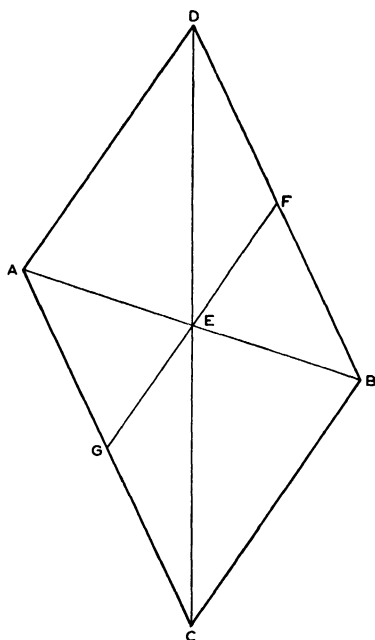


Fig. 1

Denoting now the same points with respect to f' in \mathfrak{F} by primes, we have by virtue of this construction in f and the invariance property of Theorem 4,

$$(1) \quad E' = \frac{1}{2}(C' + D'),$$

$$(2) \quad F' = \frac{1}{2}(B' + D'),$$

$$(3) \quad G' = \frac{1}{2}(A' + C'),$$

$$(4) \quad E' = \frac{1}{2}(F' + G').$$

Substituting (2) and (3) into (4) we have:

$$\begin{aligned} E' &= \frac{1}{2}[\frac{1}{2}(B' + D') + \frac{1}{2}(A' + C')] \\ &= \frac{1}{2}[\frac{1}{2}(A' + B') + \frac{1}{2}(C' + D')]. \end{aligned}$$

Now substituting (1) into the right-hand side of the last equation and simplifying, we infer the desired result:

$$E' = \frac{1}{2}(A' + B'),$$

since by construction $E = \frac{1}{2}(A + B)$.

Thus the midpoint of an arbitrary segment is preserved. Q.E.D.

THEOREM 6. *The property of two finite segments of inertial paths being parallel and in a fixed ratio is invariant in \mathfrak{F} .*

PROOF: Let $f(x) - f(y) = k[f(u) - f(v)]$, with $f(x) - f(y)$ and $f(u) - f(v)$ segments of inertial paths. Without loss of generality we may assume $k \geq 1$. Let z be the point such that $f(x) - f(y) = k[f(x) - f(z)]$. We now construct a parallelogram with $f(u) - f(v)$ and $f(x) - f(z)$ as two parallel sides. By the previous theorem any parallelogram in f is carried into a parallelogram in f' since the midpoint of the diagonals is preserved. Thus

$$(1) \quad f'(u) - f'(v) = f'(x) - f'(z),$$

but by Theorems 2 and 3

$$(2) \quad f'(x) - f'(y) = k[f'(x) - f'(z)],$$

(for details see proof of Theorem 4), whence from (1) and (2)

$$f'(x) - f'(y) = k[f'(u) - f'(v)]. \quad \text{Q.E.D.}$$

As the final theorem about properties invariant in \mathfrak{F} , we want to generalize the preceding theorem to arbitrary finite segments.

THEOREM 7. *The property of two arbitrary finite segments being parallel and in a fixed ratio is invariant in \mathfrak{F} .*

PROOF: In view of preceding theorems, the crucial thing to show is that if

$$f(x) - f(y) = k[f(x) - f(z)]$$

then

$$f'(x) - f'(y) = k[f'(x) - f'(z)].$$

Our approach is to use an "inertial" parallelogram similar to the one used in the proof of Theorem 5. In fact an exactly similar construction will be used; points A to E are constructed identically, where $A = f(x)$ and $B = f(y)$. Without loss of generality we may assume $k > 2$, that is, that $f(z) = F$ is between A and E . We then have that

$$(1) \quad A - E = \frac{k}{2}[A - F].$$

We draw through F a line parallel to CD , which cuts AC at G and AD at H . (See Figure 2.)

Now (1) is equivalent to:

$$(2) \quad F = \left(1 - \frac{2}{k}\right)A + \frac{2}{k}E.$$

Moreover, by construction

$$(3) \quad F = \frac{1}{2}(G + H)$$

$$(4) \quad E = \frac{1}{2}(C + D)$$

$$(5) \quad G = \left(1 - \frac{2}{k}\right)A + \frac{2}{k}C$$

$$(6) \quad H = \left(1 - \frac{2}{k}\right)A + \frac{2}{k}D.$$

Since GFH , AGC , AHD and CED are by construction segments of in-

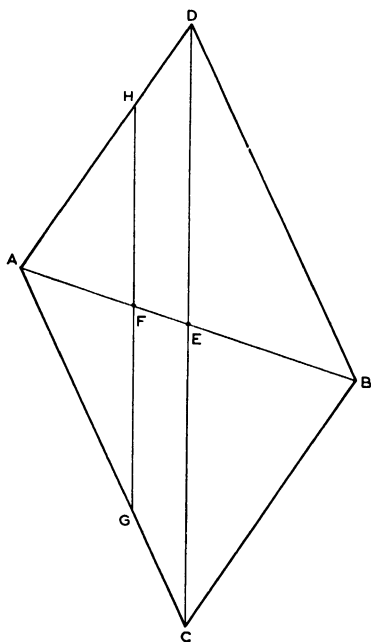


Fig. 2

But now by virtue of Theorem 5

$$E' = \frac{1}{2}(A' + B'),$$

which together with (12) yields:

$$F' = \left(1 - \frac{1}{k}\right)A' + \frac{1}{k}B',$$

which is equivalent to:

$$(13) \quad f'(x) - f'(y) = k[f'(x) - f'(z)].$$

The remainder of the proof, based upon considering $f(x) - f(y) = k[f(u) - f(v)]$, is exactly like that of Theorem 6 and may be omitted. (In place of Theorems 2 and 3 in that proof we use the result just established.)

Q.E.D.

We now state the theorem toward which the preceding seven have been directed.

ertial paths, by virtue of Theorem 7, we have from (3)–(6):

$$(7) \quad F' = \frac{1}{2}(G' + H')$$

$$(8) \quad E' = \frac{1}{2}(C' + D')$$

$$(9) \quad G' = \left(1 - \frac{2}{k}\right)A' + \frac{2}{k}C'$$

$$(10) \quad H' = \left(1 - \frac{2}{k}\right)A' + \frac{2}{k}D'.$$

Substituting (9) and (10) in (7), we get:

$$(11) \quad F' = \left(1 - \frac{2}{k}\right)A' + \frac{1}{k}(C' + D').$$

And now substituting (8) in (11), we obtain the desired result:

$$(12) \quad F' = \left(1 - \frac{2}{k}\right)A' + \frac{2}{k}E'.$$

THEOREM 8. *Any two frames in \mathfrak{F} are related by a non-singular affine transformation.*

PROOF: A familiar necessary and sufficient condition that a transformation of a vector space be affine is that parallel finite segments with a fixed ratio be carried into parallel segments with the same fixed ratio. (See, e.g. Birkhoff and MacLane [1, p. 263].) Hence by virtue of Theorem 7 any two frames are related by an affine transformation. Non-singularity of the transformation follows from the fact that each frame in \mathfrak{F} is a one-one mapping of X onto R_4 . Q.E.D.

Once we have any two frames in \mathfrak{F} related by an affine transformation, it is not difficult to proceed to show that they are related by a Lorentz transformation. In the proof of this latter fact, it is convenient to use a Lemma about Lorentz matrices, which is proved in Rubin and Suppes [3], and is simply a matter of direct computation.

LEMMA 1. *A matrix \mathcal{A} (of order 4) is a Lorentz matrix if and only if*

$$\mathcal{A} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & -c^2 \end{pmatrix} \mathcal{A}^* = \begin{pmatrix} \mathcal{I} & 0 \\ 0 & -c^2 \end{pmatrix}$$

We now prove the basic result:

THEOREM 9. *Any two frames in \mathfrak{F} are related by a Lorentz transformation.*

PROOF: Let f, f' be two frames in \mathfrak{F} . As before, for x in X , $f(x) = \mathbf{x}$, $f_1(x) = x_1$, $f'(x) = \mathbf{x}'$, etc. We consider the transformation φ such that for every x in X , $\varphi(\mathbf{x}) = \mathbf{x}'$. By virtue of Theorem 8 there is a non-singular matrix (of order 4) and a four-dimensional vector B such that for every x in X

$$\varphi(\mathbf{x}) = \mathbf{x}\mathcal{A} + B.$$

The proof reduces to showing that \mathcal{A} is a Lorentz matrix.

Let

$$(1) \quad \mathcal{A} = \begin{pmatrix} \mathcal{D} & E^* \\ F & g \end{pmatrix}.$$

And let α be a light line (in f) such that for any two distinct points \mathbf{x} and \mathbf{y}

of α if $\mathfrak{x} = \langle Z_1, t_1 \rangle$ and $\mathfrak{y} = \langle Z_2, t_2 \rangle$, then

$$(2) \quad \frac{Z_1 - Z_2}{t_1 - t_2} = W.$$

Clearly $|W| = c$. Now let

$$(3) \quad W' = \frac{Z'_1 - Z'_2}{t'_1 - t'_2}.$$

From (1), (2) and (3) we have:

$$(4) \quad W' = \frac{(Z_1 - Z_2)\mathcal{D} + (t_1 - t_2)F}{(Z_1 - Z_2)E^* + (t_1 - t_2)g}$$

Dividing all terms on the right of (4) by $t_1 - t_2$, and using (2), we obtain:

$$(5) \quad W' = \frac{W\mathcal{D} + F}{WE^* + g}.$$

At this point in the argument we need to know that $|W'| = c$, that is to say, we need to know that if $I_f(xy) = 0$, then $I_{f'}(xy) = 0$. The proof of this fact is not difficult. From our fundamental invariance axiom we have that $I_f(xy) \geq 0$, that is,

$$(6) \quad |W'| \geq c.$$

Consider now a sequence of inertial lines $\alpha_1, \alpha_2, \dots$ whose slopes W_1, W_2, \dots are such that

$$(7) \quad \lim_{n \rightarrow \infty} W_n = W.$$

Now corresponding to (5) we have:

$$(8) \quad |W'_n| = \left| \frac{W_n\mathcal{D} + F}{W_nE^* + g} \right| < c.$$

Whence, from (8) we conclude that if $WE^* + g \neq 0$, then

$$(9) \quad |W'| = |\lim_{n \rightarrow \infty} W'_n| \leq c.$$

Thus from (6) and (9) we infer

$$(10) \quad |W'| = c,$$

if $WE^* + g \neq 0$, but that this is so is easily seen. For, suppose not. Then

$$\lim_{n \rightarrow \infty} (W_n E^* + g) = 0,$$

and thus

$$\lim_{n \rightarrow \infty} (W_n \mathcal{D} + F) = 0.$$

Consequently $W\mathcal{D} + F = 0$, and $\langle W, 1 \rangle \mathcal{A} = 0$, which is absurd in view of the non-singularity of \mathcal{A} .

Since $|W'| = c$, we have by squaring (5):

$$(11) \quad \frac{W\mathcal{D}\mathcal{D}^*W^* + 2W\mathcal{D}F^* + |F|^2}{(WE^* + g)^2} = c^2,$$

and consequently

$$(12) \quad W(\mathcal{D}\mathcal{D}^* - c^2E^*E)W^* + 2W(\mathcal{D}F^* - c^2E^*g) + |F|^2 - c^2g = 0.$$

Since (12) holds for an arbitrary light line, we may replace W by $-W$, and obtain (12) again. We thus infer:

$$W(\mathcal{D}F^* - c^2E^*g) = 0,$$

but the direction of W is arbitrary, whence

$$(13) \quad \mathcal{D}F^* - c^2E^*g = 0.$$

Now let $\mathbf{x} = \langle 0, 0, 0, 0 \rangle$ and $\mathbf{y} = \langle 0, 0, 0, 1 \rangle$. Then

$$I_j^2(xy) = -c^2.$$

But it is easily seen from (1) that

$$I_j^2(xy) = |F|^2 - c^2g^2,$$

and thus by our fundamental invariance axiom

$$(14) \quad c^2g^2 - |F|^2 = c^2.$$

From (12), (13), (14) and the fact that $|W|^2 = c^2$, we infer:

$$W(\mathcal{D}\mathcal{D}^* - c^2E^*E)W^* = |W|^2,$$

and because the direction of W is arbitrary we conclude:

$$(15) \quad \mathcal{D}\mathcal{D}^* - c^2E^*E = \mathcal{I},$$

where \mathcal{I} is the identity matrix.

Now by direct computation on the basis of (1),

$$(16) \quad \mathcal{A} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & -c^2 \end{pmatrix} \mathcal{A}^* = \begin{pmatrix} \mathcal{D}\mathcal{D}^* - c^2 E^* E & \mathcal{D}F^* - c^2 E^* g \\ (\mathcal{D}F^* - c^2 E^* g)^* & FF^* - c^2 g^2 \end{pmatrix}$$

From (13), (14), (15) and (16) we arrive finally at the result:

$$\mathcal{A} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & -c^2 \end{pmatrix} \mathcal{A}^* = \begin{pmatrix} \mathcal{I} & 0 \\ 0 & -c^2 \end{pmatrix},$$

and thus by virtue of Lemma 1, \mathcal{A} is a Lorentz matrix. Q.E.D.

4. Temporal Parity. Turning now to problems of parity, we may for simplicity restrict the discussion to time reversals. Similar considerations apply to spatial reflections.

A simple axiom, which will prevent time reversal between frames in \mathfrak{F} , is:

(T1) *There are elements x and y in X such that for all f in \mathfrak{F}*

$$f_4(x) < f_4(y).$$

There is, however, a simple objection to this axiom. It is unsatisfactory to have time reversal depend on the existence of special space-time points, which could possibly occur only in some remote region or epoch. This objection is met by T2.

(T2) *If $I_1^2(xy) < 0$ then either for all f in \mathfrak{F}*

$$f_4(x) < f_4(y)$$

or for all f in \mathfrak{F}

$$f_4(y) < f_4(x).$$

T2 replaces the postulation of special points by a general property: given any segment of an inertial path, all frames in \mathfrak{F} must orient the direction of time for this segment in the same way.

Nevertheless, there is another objection to T1 which holds also for T2: the appropriate axiom should be formulated so that a given observer in a frame f may verify it without observing any other frames, that is, he may decide if he is a qualified candidate for membership in \mathfrak{F} without observing other members of \mathfrak{F} . (This issue is relevant to the single axiom of Definition 3 but cannot be entered into here.) From a logical standpoint this means eliminating quantification over elements of \mathfrak{F} , which may be

done by introducing a fourth primitive notion, a binary relation σ of *signaling* on X . To block time reversal we need postulate but two properties of σ :

(T3.1) *For every x in X there is a y in X such that $x\sigma y$.*

(T3.2) *If $x\sigma y$ then $f_4(x) < f_4(y)$.*

However, a third objection to (T1) also applies to (T2) and (T3). Namely, we are essentially postulating what we want to prove. The axioms stated here correspond to postulating artificially in a theory of measurement of mass that a certain object must be assigned the mass of one. I pose the question: *Is it possible to find "natural" axioms which fix a direction of time?* It may be mentioned that Robb's meticulous axiomatization [2] in terms of the notion of *after* provides no answer.

References

- [1] BIRKHOFF, G. and S. MACLANE, *A Survey of Modern Algebra*. New York 1941, XI+ 450 pp.
- [2] ROBB, A. A., *Geometry of Space and Time*. Cambridge 1936, VII + 408 pp.
- [3] RUBIN, H. and P. SUPPES, *Transformations of systems of relativistic particle mechanics*. Pacific Journal of Mathematics, vol. 4 (1954), pp. 563-601.

AXIOMS FOR COSMOLOGY

A. G. WALKER

University of Liverpool, Liverpool, England

1. In relativistic cosmology there is a generally accepted form of space-time which serves as a geometrical model for the large scale features of the universe. It is a four dimensional manifold with a quadratic differential metric

$$dt^2 - R^2 d\sigma^2$$

where t is a preferential coordinate, R is a function of t only, and $d\sigma^2$ is the metric of a three dimensional Riemannian space C of constant curvature k . Topologically the space-time is a product $T \times C$ where T is the continuum of real numbers (parametrised by t) and C is a 3-space which may be spherical or elliptic ($k > 0$), hyperbolic ($k < 0$) or euclidean ($k = 0$). Each point x of C represents a *fundamental particle* (corresponding to a galaxy in the universe); the curve $T \times x$, an orthogonal trajectory of the hypersurfaces $t = \text{constant}$ in space-time, is the world line of the particle, and the null geodesics of spacetime represent light paths. The natural projection of each such null geodesic into C is a geodesic of C .

In dynamical theories, such as General Relativity, the form of space-time is strictly invariant and the function R is significant in that, through the field equations, it determines the distribution of matter in the universe. In kinematical theories, however, space-time is only conformally invariant with the result that R can be 'transformed away' by a regraduation of the time scale, i.e. a transformation of the time parameter from t to τ where $d\tau = dt/R$. Ignoring a conformal factor, the metric of space-time then becomes $d\tau^2 - d\sigma^2$ and the model is static. The τ -scale giving this comparatively simple model is unique except for an arbitrary affine transformation $\tau' = a\tau + b$ ($a > 0$); with each τ -scale there is a 'natural' measure of distance in C , by $\int d\sigma$, and the only effect of a regraduation $\tau' = a\tau + b$ is a change of distance unit by a factor a . With these measurements of time and distance light can be said to have unit speed in C .

The above model, which is common to many theories, was derived first by Lemaitre and others from Einstein's General Theory, when it was interpreted as a dynamical as well as a kinematical model. Later it was

derived by Robertson and the writer independently as a purely kinematical model, based on fewer assumptions than in General Relativity, and this derivation is generally regarded as satisfactory and adequate in modern cosmological theories. Nevertheless it is far from satisfactory as an example of the axiomatic method largely because of the initial assumption that events can be described by numerical parameters, i.e. that the natural topology of a geometrical model of the universe is that of a manifold. This is a good working hypothesis in that it produces useful results quickly, but we now wish to base the structure on a more elementary set of axioms.

We shall be talking about *particles* (fundamental particles) and the *events* in the history of a particle, and the present purpose is to find a system of axioms from which we can deduce two theorems; firstly, that the events in the history of a single particle are 'linearly' ordered, i.e. can be parametrised by a single real parameter; secondly, that the set of particles can be given the topology and structure of a geodesic metric space in such a way that the metric has the properties of metric in the cosmological model. We shall not go all the way in establishing the spherical, elliptic, hyperbolic or euclidean manifold structures on the set of particles, but the final stage is not difficult once the geodesic metric structure is established, using the work of Busemann, Montgomery and Zippin and postulating sufficient symmetry about each particle. Our axiomatic system will in fact cut out the spherical and elliptic models since it will be postulated that the (light) signal correspondence from one particle to another is one-one. It would need a more complicated system to include the models of positive curvature and this is not discussed here.

2. Before the axioms are stated the idea of light signals used to such good affect by E. A. Milne [2] and others will be described briefly. One of the primitives in the present system, the signal-mapping of one particle set of events (world-line) on another, is based on this idea, and one of the axioms appears artificial until it is related to the situation of equivalence between particle-observers discussed by Milne.

Milne's particle-observer is the set of events in the history of a particle together with a 'clock', i.e. a numerical parameter giving temporal order in the set. If A and B are two particle-observers, light signals can be sent from one to the other; the time of arrival s' at B can be expressed as a function $s' = \theta(t)$ of the time t of emission at A , and the time of arrival t' at A is a function $t' = \bar{\theta}(s)$ of the time s of emission at B . These 'times' at

A and B are recorded by the clocks attached to A and B . The particle-observers (with their clocks) are said to be *equivalent* if the functions θ and $\bar{\theta}$, called signal functions, are identical, and it can be shown that if A and B are not equivalent, then B 's clock can be regraduated by a transformation of the form $s' = \psi(s)$ so that they become equivalent.

Three particle-observers A, B, C are collinear, with B between A and C , if the light signal from A to C is the same as that from A to B followed by the signal from B to C , and similarly from C to A . They then form an equivalent system if they are equivalent in pairs, and it is easily verified that if θ and ϕ are the signal functions between A and B and between A and C respectively, the condition for this is $\theta \circ \phi = \phi \circ \theta$. A collinear set of particle-observers equivalent in pairs thus gives rise to a set of commutative signal functions, and from the study of such a set Milne was able to establish theorems on linear equivalences.

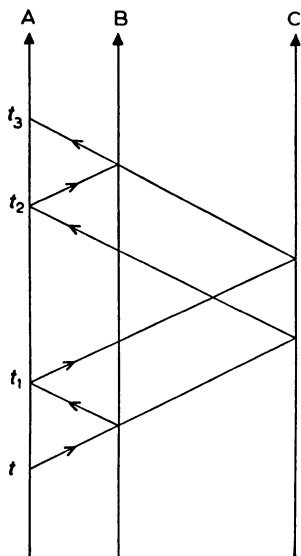


Fig. 1

One serious disadvantage of this treatment is the assumption that particle-observers can communicate with each other so that a signal function is assumed to be knowable. It would be difficult to embody this assumption in an axiomatic system and for that reason it will be assumed in the present work that all observations are to be made by only *one* observer. Thus if A is this observer and if A and B are equivalent in Milne's sense, A cannot observe the signal function θ between A and B but he can observe the function $\theta^2 = \theta \circ \theta$, since if a light signal emitted by A at time t is reflected at B and is received by A at time t' , then t' is an observable function of t given by $t' = \theta^2(t)$.

Again, if collinear particles A, B, C are equivalent in Milne's sense and if θ, ϕ are the signal functions between A and B and between A and C as before, then $\theta \circ \phi = \phi \circ \theta$; but if A is the only observer, this is not an observable relation. A consequence, however, is $\theta^2 \circ \phi^2 = \phi^2 \circ \theta^2$, and this is an observable relation since θ^2 and ϕ^2 are observable. This simple relation is independent of the choice of clock scales and can be illustrated as in fig. 1, where the

vertical lines represent world-lines and the other lines light paths. Although derived from Milne's idea of equivalence it is in fact weaker, and provides the main suggestion for our Axiom IX which is equivalence to it when applied to collinear particles.

3. The primitives of the axiomatic system to be considered here are *events* and certain sets of events called *particles*. The events of one particle O , called the *observer*, satisfy a *total order relation*, described by the words 'before' and 'after'; if x and y are distinct events of O , then either $x < y$ (x is before y , equivalent to $y > x$, i.e. y is after x) or $y < x$. Lastly, if A and B are any two particles, there is a *signal-mapping* of A onto B , denoted by (A, B) .

AXIOM I. *The order relation in O is transitive, i.e. if x, y, z , are events of O such that $x < y$ and $y < z$, then $x < z$.*

AXIOM II. *Every signal-mapping is one-one.*

Thus (A, B) has a single valued inverse $(A, B)^{-1}$ which is a mapping of B onto A .

DEFINITION. *An OBSERVABLE is a mapping $O \rightarrow O$ resulting from a chain of signal mappings or inverse signal mappings.*

One example of an observable is the signal-mapping (O, A) followed by the mapping (A, O) ; this will be denoted by $(O, A)(A, O)$, which is here more convenient than the usual $(A, O) \circ (O, A)$. Another example is $(O, A)(A, B)(O, B)^{-1}$, which will turn out later to be the identity mapping $O \rightarrow O$ when O, A, B are 'collinear'.

All further axioms can now be expressed in terms of observables and the order relation on O , but for convenience we shall define and use *relative observables*.

By means of the mapping (O, A) and the order relation on O , an order relation can be induced on A , and we shall use the symbols $<, >$ and words 'before' and 'after' when describing this relation. An OBSERVABLE RELATIVE TO A is defined as a mapping $A \rightarrow A$ resulting from a chain of signal mappings and inverses, and an axiom may be expressed in terms of observables relative to any particle A and the order relation on A . This is only a matter of convenience; any such expression could always be restated in terms of proper observables, i.e. observables relative to O , for if f is an observable relative to A , then $(O, A) f (O, A)^{-1}$ is a corresponding proper observable.

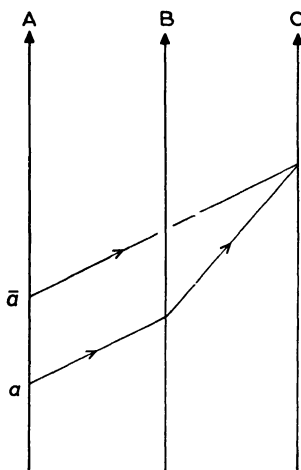


Fig. 2

Let A, B, C be any three particles, and let g be the observable relative to A defined as follows (see fig. 2)

$$g = (A, B)(B, C)(A, C)^{-1}.$$

AXIOM III. $g(a) \geq a$ for all events $a \in A$.

AXIOM IV. g is strictly increasing, i.e. if a, a' are events of A such that $a' > a$, then $g(a') > g(a)$.

It is to be understood that in axioms such as these the particles are not necessarily distinct, i.e. B or C may be a copy of A , with the convention that the signal mapping (A, A) is the identity $A \rightarrow A$. Putting $C = A$ in Axioms III and IV we thus have that the observable f relative to

A , defined by $f = (A, B)(B, A)$, satisfies the same conditions as g in the axioms.

We also see from Axiom IV with $A = O$ that if B and C are any two particles, the order relation induced in C from B by means of the mapping (B, C) is the same as the order induced in C from O . It follows that the observer O loses its preferential position; the whole system is the same relative to a 'subordinate observer' at A with the order relation induced from O .

It can now be assumed without a further axiom that the ordered set O is closed, and therefore that every particle set is closed, i.e. that every bounded sequence of events in a particle has a limit. If the particle sets are not closed, new events can be defined in the usual way as sections or by sequences, and the sets of new events are closed. Further, the signal mappings can be extended in a natural way to the new particle-sets so that the above axioms are still satisfied. It will therefore be assumed that O , and hence every particle set of events, is closed.

DEFINITION. *Particles A and B COINCIDE at the event $a \in A$ if $f(a) = a$ where $f = (A, B)(B, A)$.*

We see that if A and B coincide at $a \in A$ and if $(A, B)(a) = b$, then A and B coincide at the event $b \in B$; for we have $f(a) = (B, A) \circ (A, B)(a) = (B, A)(b)$ and hence $(B, A)(A, B)(b) = (A, B) \circ (B, A)b = (A, B)(a) = b$.

We could, if we wished to follow the mechanical picture, regard a and b here as the same event and so allow particle sets to intersect. This is unnecessary, however, because our particles are restricted to correspond to what were formerly called fundamental particles and therefore are required not to coincide, which leads to the next axiom.

AXIOM V. *No two distinct particles coincide at any event.*¹

It follows that no particle set has a first or last event, (assuming that there is more than one particle; see Axiom VII)². For if a is a last event of a particle A and B is another particle, then by Axiom III with $C = A$, $f(a) \geq a$ where $f = (A, B)(B, A)$ and hence $f(a) = a$, i.e. A and B coincide at a . Similarly A and B would coincide at a first event of A .

4. DEFINITION. *Particles A, B, C are COLLINEAR, with B between A and C , if*

$$(A, B)(B, C) = (A, C), \quad (C, B)(B, A) = (C, A).$$

These conditions can be expressed in terms of observables and requires the observables relative to A , given by $(A, B)(B, C)(A, C)^{-1}$ and $(C, A)^{-1}(C, B)(B, A)$, both to be the identity mapping $A \rightarrow A$. They correspond, therefore, to the case of equality in Axiom III.

AXIOM VI. *If A, B, C, D are particles with A, B, C collinear in some order, A, B, D collinear in some order, and A, B distinct, then A, C, D are collinear in some order.*

It follows from this that the set of all particles collinear, in some order, with two distinct particles is a *linear system* which is determined by any two distinct members. From the 'between' relation in the above definition and Axiom III we see that a linear system is totally ordered. In particular we can talk about particles of a linear system being on the 'same side' or on 'opposite sides' of a member of the system.

¹ This axiom is in fact redundant, but to do without it would mean a great deal of additional work. A theorem equivalent to this axiom is proved in [4] where also Axiom VIII is weakened.

² We could, of course, make an exception of first and last events in Axiom V, and it would then follow that if one particle has, for example, a first event then all particles coincide at this event, which is what happens in Milne's model with the t -scale of time. However, extreme events of this kind can be excluded without affecting the system and the form of Axiom V given here appears to be preferable.

DEFINITION. A linear system L is DENSE at a particle $A \in L$ if, for any two events a, a' of A with $a' > a$, there is a particle $B \in L$ distinct from A such that $(A, B)(B, A)(a) < a'$.

It is not difficult to prove that, if L is dense at A , there is a sequence $\{A_n\}$ of particles in L and on the same side of A such that, for every event $a \in A$, $f_n(a) \rightarrow a$ as $n \rightarrow \infty$ where $f_n = (A, A_n)(A_n, A)$. We shall write $A_n \rightarrow A$.

It should be noted that this definition of denseness in a linear system is stronger than denseness in the ordinary sense for a totally ordered set since it involves the ordered set A of events. One consequence, of course, is that L is dense at A in the ordinary sense that if B, C are members of L with A between them, then there is another member of L which is distinct from A, B, C and between B and C . Another consequence is as follows.

If there is a linear system of particles which is dense at some member then the particle set O of events (and hence every particle set) is *continuous* in the sense that if x, y are any two events of O and $x < y$, there is an event z of O such that $x < z < y$.

AXIOM VII. *There are at least two distinct particles.*

Hence there is at least one linear system of particles.

AXIOM VIII. *Every linear system of particles is dense at every member*³.

An immediate consequence of this axiom, as remarked above, is that every particle set of events is continuous. We are now in a position to prove the following theorem.

The particle set O , and hence every particle set, is ordinally equivalent to the continuum of real numbers.

It is sufficient to prove this for any one particle A , and since we already have that the set A is closed, it is sufficient to prove that there is an enumerable subset of A such that between any two events of A is a member of the subset.

Let L be a linear system containing A . Then L is dense at A by Axiom VIII and there is a subset $\{A_n\}$ of L such that $A_n \rightarrow A$. As before we write f_n for $(A, A_n)(A_n, A)$ and define f_n^p for any integer p in the usual

³ Because of an axiom of symmetry which comes later (§ 6), Axiom VIII could be weakened to state that every linear system is dense at some member; it would then follow from symmetry that the system is dense at every member. The present axiom is chosen, however, so that certain theorems can be proved immediately. (Cf. [4].)

way; thus f_n^0 is the identity mapping $A \rightarrow A$, $f_n^1 = f_n$, $f_n^{p+1} = f_n \circ f_n^p$, and $f_n^{-p} = (f_n^p)^{-1}$.

Let a be an event of A , and consider the subset of A given by the events $f_n^p(a)$ where n takes all positive integer values and p takes all integer values. This subset is enumerable, and we shall prove that if x, y are any two events of A and $x < y$, there is a member of this subset between x and y . It will be sufficient to consider the case $a < x < y$; the proof for the case $x < y < a$ is very similar, and the case $x < a < y$ is trivial.

Let x, y be events of A and $a < x < y$. Then since $A_n \rightarrow A$, there is an n such that $f_n(x) < y$. Keeping n fixed, consider the sequence $f_n^m(a)$ where m takes positive or zero integer values. This sequence is unbounded as $m \rightarrow \infty$, for if $f_n^m(a) \rightarrow z$ as $m \rightarrow \infty$ then $f_n(z) = z$ and A_n coincides with A at the event z , which contradicts Axiom V. Hence, since $a < x$, there is a positive or zero integer m such that

$$f_n^m(a) \leq x < f_n^{m+1}(a).$$

The mapping f is increasing by Axiom IV and hence

$$x < f_n^{m+1}(a) = f_n(f_n^m(a)) \leq f_n(x) < y$$

i.e. the event $f_n^{m+1}(a)$ in the enumerable subset of A is between x and y , as required.

The proof for $x < y < a$ is similar but with inverse signal mappings f_n^{-m} .

This theorem shows that every particle set can be mapped onto the continuum of real numbers so that order is preserved in the sense that 'before' corresponds to 'is less than'. Such a mapping will be called a *clock*, and the real number corresponding to an event is a *clock reading*. The mapping is not, of course, unique, and a change of mapping may be called a 'clock regraduation'; it corresponds to a transformation $t' = \psi(t)$ of the 'time' parameter t , where ψ is a continuous increasing function taking all values.

When a particle A is provided with a clock in this way, every observable relative to A can be represented as a function of the time parameter, and from the axioms it follows that all such functions are continuous and increasing and take all values. If f is such a function, it is transformed into $\psi \circ f \circ \psi^{-1}$ when A 's clock is regraduated by $t' = \psi(t)$.

5. The next axiom was suggested by a property of Milne's equivalent system of collinear particle-observers (see § 2) but applies to any three particles, not necessarily collinear.

AXIOM IX. *If A, B, C are any three particles then $(A : B, C) = (A : C, B)$ where $(A : B, C)$ denotes the observable $(A, B)(B, C)(C, B)(A, B)^{-1}$ relative to A .*

If A, B, C are collinear with A not between B and C this axiom is seen to be already satisfied. If however A is between B and C , then $(C, B)(A, B)^{-1} = (C, A)$ and the axiom gives

$$(A, B)(B, C)(C, A) = (A, C)(C, B)(B, A)$$

i.e.

$$(A, B)(B, A)(A, C)(C, A) = (A, C)(C, A)(A, B)(B, A),$$

showing that the two observables $(A, B)(B, A)$ and $(A, C)(C, A)$ relative to A commute.

Again, if A, B, C are collinear with B between A and C , then the axiom applied to B, A, C in this order gives

$$(B, A)(A, C)(C, B) = (B, C)(C, A)(A, B)$$

and hence

$$(A, B)(B, A)(A, C)(C, B)(B, A) = (A, B)(B, C)(C, A)(A, B)(B, A)$$

i.e.

$$(A, B)(B, A)(A, C)(C, A) = (A, C)(C, A)(A, B)(B, A)$$

since $(C, B)(B, A) = (C, A)$ and $(A, B)(B, C) = (A, C)$. Thus the observables $(A, B)(B, A)$ and $(A, C)(C, A)$ relative to A commute as before. A similar result occurs if C is between A and B . Hence:

If A, B, C are collinear in some order, the observables $(A, B)(B, A)$ and $(A, C)(C, A)$ relative to particle A commute.

Consider now a linear system L of particles containing A , and for $X \in L$ denote by f_X the observable $(A, X)(X, A)$ relative to A . Then from what has just been proved, if X, Y are any two members of L ,

$$f_X \circ f_Y = f_Y \circ f_X.$$

Suppose now A is assigned a clock, i.e. a 'time' parameter t ; then observables such as f_X are represented by continuous increasing functions $f_X(t)$ taking all values, and any two such functions corresponding to members of L commute. We thus have a system L of commutative functions which, because of the denseness of L at A , contains a sequence which converges uniformly to the identity. Hence, from a theorem on sets

of commutative functions [3], there exists a continuous increasing function $\psi(t)$, taking all values, such that every function $f_X \in L$ can be expressed in the form.

$$f_X(t) = \psi^{-1}\{2d_X + \psi(t)\}$$

where d_X is a positive or zero constant depending upon X .

If now A 's clock is regraduated to read time τ where $\tau = \psi(t)$, the observable function $f_X(t)$ is transformed into the function $\tau + 2d_X$. We have thus proved that:

If L is a linear system of particles containing A , a clock reading time τ can always be assigned to A so that, if X is any member of L , the observable $(A, X)(X, A)$ relative to A is represented by the function $\tau + 2d_X$ where d_X is a positive or zero constant depending upon X . Such a clock will be called a τ -CLOCK relative to L .

It also follows from the theorems on commutative functions that A 's τ -clock is determined uniquely by the linear system except for an arbitrary affine regraduation $\tau' = a\tau + b$, $a > 0$. The only effect of such a regraduation on the observable functions $\tau + 2d_X$ is to multiply all the constants d_X by the same factor a .

We now define the DISTANCE $d(A, X)$ from A to X to be d_X . We observe that $d(A, X) \geq 0$, and equality occurs when and only when $X = A$. In terms of readings on A 's τ -clock the distance $d(A, X)$ is given by

$$d(A, X) = \frac{1}{2}(\bar{\tau} - \tau)$$

where τ has any value and $\bar{\tau} = f_X(\tau)$, $f_X = (A, X)(X, A)$. This formula indicates again how the 'scale' of distance depends upon the choice of τ -clock; under the allowable change of τ -scale given by $\tau' = a\tau + b$ we get

$$d'(A, X) = \frac{1}{2}(\bar{\tau}' - \tau') = \frac{1}{2}a(\bar{\tau} - \tau) = ad(A, X).$$

When a τ -clock has been assigned to A , a clock can be assigned to any other particle $X \in L$ by taking the clock reading at the event $(A, X)(\tau)$ to be $\tau + d_X$, and it can easily be verified that this parametrisation of X is for X a proper τ -clock relative to L .

DEFINITION. *The τ -clock assigned to X as above is EQUIVALENT to A 's τ -clock.*

It can be verified that for all particles of L and τ -clocks relative to L , equivalence as defined here is reflexive and transitive. Also, for any

particles X, Y of L , the distance $d(X, Y)$ measured in relation to a τ -clock of X (relative to L) is equal to the distance $d(Y, X)$ measured in relation to the equivalent τ -clock of Y , and for any three particles X, Y, Z of L , with Y between X and Z ,

$$d(X, Z) = d(X, Y) + d(Y, Z)$$

where distances are measured in relation to equivalent τ -clocks relative to L .

If a τ -clock undergoes an additive regraduation $\tau' = \tau + b$, it becomes a τ -clock and distances measured in relation to it are unaltered. However, if one of two equivalent τ -clocks undergoes this regraduation with $b \neq 0$, it ceases to be equivalent to the other, and we shall say that the clocks are then *congruent*.

DEFINITION. *If τ -clocks relative to L are attached to two particles of a linear system L and are equivalent to within additive regraduations, they are CONGRUENT.*

From the properties of equivalence it follows that for all particles of L and τ -clocks relative to L , congruence is reflexive and transitive. Also, the distance relations $d(X, Y) = d(Y, X)$, $d(X, Z) = d(X, Y) + d(Y, Z)$ for particles of L hold when distances are measured in relation to congruent τ -clocks.

6. We now wish to extend this idea of distance from a linear system to the whole system of particles and so establish a metric on the 'space' of particles. For this we need the general form of Axiom IX together with a new axiom of symmetry.

Consider first a mapping ρ of the set of particles onto itself which leaves a particle A invariant. An observable f relative to A is a mapping $A \rightarrow A$ determined in some way by a sequence of particles B, C, \dots , and the transform of f under ρ may be defined as the observable $A \rightarrow A$ determined in the same way by the sequence B', C', \dots where $B' = \rho(B)$, etc.

DEFINITION. *An A-TRANSFORMATION is a one-one mapping of the set of particles onto itself which leaves A invariant and is such that every observable relative to A is transformed into itself.*

Since the property of collinearity of particles can be defined in terms of observables relative to any particle A , it follows that a linear system of particles is mapped onto a linear system by any A -transformation. Also,

if L is a linear system containing A and if L is mapped onto L' by an A -transformation, a τ -clock of A relative to L is also a τ -clock relative to L' . If X, Y are members of L and $d(X, Y)$ the distance between them in relation to a τ -clock of A , and if X', Y' are the images of X, Y under the A -transformation and $d(X', Y')$ the distance between them in relation to the same τ -clock of A , then $d(X', Y') = d(X, Y)$.

DEFINITION. *A HALF-LINE at a particle A is part of a linear system containing A ; it consists of A and all the particles on one side of A .*

Thus a linear system containing A is the union of two half-lines at A . If B is a particle distinct from A , there is just one half-line at A which contains B .

AXIOM X. *If A is any particle and l, m any two half-lines at A , there is an A -transformation which maps l onto m .*

We can now talk of any observer A being assigned a τ -clock without reference to any particular linear system containing A ; a τ -clock relative to one such linear system will also be a τ -clock relative to any other because of Axiom X and the property of an A -transformation mentioned above. Thus to every observer can be assigned a τ -clock which is unique to within an arbitrary affine regratuation.

The definition of congruence of τ -clocks given in § 5 applies to any two observers, and we can now prove that this congruence is transitive, i.e. if to any three observers, A, B, C are assigned τ -clocks such that those of A and B are congruent and those of A and C are congruent, then those of B and C are congruent. To prove this it is sufficient to prove that the distances $d(B, C)$ and $d(C, B)$, measured in relation to the τ -clocks assigned to B and C respectively, are equal. This is a consequence of the formula $\frac{1}{2}(\bar{\tau} - \tau)$ for distance in terms of τ -clock readings and the definition of congruent τ -clocks applied to the pairs A, B and A, C , for we find that, for any number τ ,

$$d(B, C) = \frac{1}{2}\{(A : A, C)(\tau) - \tau\}, \quad d(C, B) = \frac{1}{2}\{(A : A, B)(\tau) - \tau\}$$

and these are equal by Axiom IX.

Thus τ -clocks can be assigned to all particles so that they are congruent in pairs, and if a particular τ -clock is assigned to one particle, say O , the congruent τ -clock attached to any other particle is unique to within an arbitrary additive regratuation $\tau' = \tau + b$. Such a regratuation does not affect measurements of distance and hence the distance $d(X, Y)$ between

any two particles X, Y is uniquely determined. If O 's τ -clock undergoes an affine regratuation $\tau' = a\tau + b$, $a > 0$, then all distances are multiplied by the same factor a . We have thus defined a METRIC on the set of particles, and it is easily verified that all the fundamental properties of a metric are satisfied. For example, the triangular inequality

$$d(A, C) \leq d(A, B) + d(B, C)$$

is mainly a consequence of Axiom III, for we have in terms of A 's τ -clock and for any number τ , $2d(B, C) = \bar{\tau} - \tau$ where

$$\bar{\tau} = (A, B)(B, C)(C, B)(A, B)^{-1}(\tau)$$

Also,

$$2d(A, B) = (A, B)(B, A)(\bar{\tau}) - \bar{\tau}$$

and hence

$$2d(A, B) + 2d(B, C) = (A, B)(B, C)(C, B)(B, A)(\tau) - \tau.$$

Since by Axiom III, $(A, B)(B, C)(\tau) \geq (A, C)(\tau) = \tau'$ say, and $(C, B)(B, A)(\tau') \geq (C, A)(\tau')$ we have from Axiom IV

$$2d(A, B) + 2d(B, C) \geq (C, B)(B, A)(\tau') - \tau \geq (C, A)(\tau') - \tau$$

i.e. $2d(A, B) + 2d(B, C) \geq (A, C)(C, A)(\tau) - \tau = 2d(A, C)$

as required.

The structures we have given to the set of particles is not merely that of a metric space; it is that of a *geodesic metric space* as defined by Busemann [1], for it is easily verified that the linear systems of particles are geodesics of the metric space and have the properties required for a geodesic space. We have thus reached our second objective.

7. By axiom X the geodesic space of particles is symmetric in the sense that for any particle A and two half-geodesics at A , there is an isometry of the space which leaves A invariant and maps one half-geodesic on the other. From what is already known about geodesic spaces it would not be difficult to select further axioms to ensure that the space is 3-dimensional hyperbolic or euclidean. For example, we could define a ROTATION about A as an isometry which is either the identity mapping or leaves one and only one geodesic through A point-wise invariant; then replace ' A -transformation' by 'rotation' in Axiom X and postulate that the set of all rotations about A is a group.

The final task in the derivation of the space-time model is to establish τ as a 'cosmic' coordinate, i.e. to show that all the particles can be assigned τ -clocks which are not merely congruent but also equivalent to each other. We then have the product structure $T \times C$ on the set of all events, C being the space of particles, and it is a straightforward matter to define a metric on $T \times C$, determine light paths (defined in terms of linear systems and signal mappings) and so complete the features of the cosmological model.

Equivalent τ -clocks have already been defined and it is an open question whether the transitivity of this equivalence for all particles is a consequence of the axioms already given. If necessary it is a simple matter to find an additional axiom which gives the property of transitivity. For example, if A, B, C are any three particles with τ -clocks such that those of A and B are equivalent and those of A and C are equivalent, it can be verified that the observable functions $(A, B)(B, C)(C, A)(\tau)$ and $(A, C)(C, B)(B, A)(\tau)$ are both of the form $\tau + \text{constant}$, and that they are the same function if and only if the τ -clocks of B and C are equivalent. It would be sufficient, therefore to postulate:

AXIOM XI. *If A, B, C are any particles,*

$$(A, B)(B, C)(C, A) = (A, C)(C, B)(B, A).$$

It is possible, however, that this is a consequence of the previous axioms.

Bibliography

- [1] BUSEMANN, H., *The Geometry of Geodesics*. Academic Press Inc. New York, 1955, X + 422 pp.
- [2] MILNE, E. A., *Kinematic Relativity*. Oxford 1948, VII + 238 pp.
- [3] WALKER, A. G., *Commutative functions*, I. Quarterly Journal of Mathematics vol. 17 (1946) pp. 65-82.
- [4] ———, *Foundations of Relativity*. Proceedings of the Royal Society of Edinburgh. vol. 62 (1948) pp. 319-335.

**AXIOMATIC METHOD AND THEORY OF RELATIVITY
EQUIVALENT OBSERVERS AND
SPECIAL PRINCIPLE OF RELATIVITY**

YOSHIO UENO

Hiroshima University, Hiroshima, Japan

1. **Axiomatization of Relativity Theory.** Roughly, speaking there are two different approaches when we try to examine the foundation of relativity by means of axiomatic methods. In the first approach one tries to axiomatize the theory of relativity as it is now. According to the second, one does not necessarily aim at deriving the present theory. Rather, one investigates various possible ways of axiomatizing the theory of relativity, in the hope that one will be able to examine prospective forms of new theories.

In the first approach, one postulates at the beginning the present relativity theory as the firmly established theory and asks what set of axioms is equivalent to the theory. Most of the works done so far has taken this approach. Certainly, most people accept general relativity as well as special relativity as firmly established theories, just like classical mechanics and electrodynamics.

However, one needs to reinvestigate some of the fundamental concepts of relativity such as space-time, scale, clock and equivalence of observers, although they are now regarded as completely established beyond any doubt. For instance, the fact that the so-called clock paradox is still discussed today indicates that there remains some ambiguity about the definition and interpretation of an observer or a moving clock.

Furthermore, we know some examples of peculiar structure of space-time as shown by Gödel's peculiar cosmological solution [1] and also by another peculiar solution due to Nariai [2]. We cannot reject these peculiar solutions only from fundamental principles of relativity. This may be again a reason for reinvestigating fundamental principles of relativity. Of course, to these peculiar solutions, the respective authors gave physical interpretations which seem reasonable. However, to insure the validity of such interpretations, we will have to understand clearly the fundamental principles of general relativity. It is beyond any doubt that axiomatic

methods are very useful for the study of this kind. I will not, however, go into details of such studies here.

Comparing these two alternative approaches, we may say that while logical formulation is the central problem of the first, heuristic considerations play the main part in the second. Namely, according to the latter viewpoint the main subject will be to examine in what forms one can formulate the fundamental concepts of relativity.

From now on I want to deal with the second approach of axiomatic formulations, namely, how to formulate physical principles of relativity. In this approach, we are not anticipating the reproduction of special and general relativity in their present form and content. Rather, my main concern will be how one can possibly change their content.

Then, what would be the fundamental concept that I should examine first? One may start from considering the relation between matter and space-time. Or one may consider first observers and invariance of physical laws. The latter was the main subject of the work on equivalent observers, which I did with Takeno [3], and also of my work [4] on equivalent observers in special relativity. I shall deal mainly with the subject of observers and their equivalence. Most of the content of this paper is from the papers I just mentioned.

2. Equivalent Observers. In general relativity, matter and space-time are specified by each other, and this is one of the basic characteristics of the theory. In special relativity matter does not affect directly the structure of space-time. There, the space-time is independent of the presence of matter and is an external element which defines modes of existence of physical phenomena. In special relativity, such modes of existence of physical phenomena are determined in reference to the state of an observer.

It is for this reason that we brought up the concept of observers as the starting point of our work. We considered first the existence of an observer and discussed its kinematical aspect. Following the work by Takeno and Ueno [3], I will explain how this was actually done. The first postulate we made was the existence of a three dimensional space frame and a one dimensional time frame for an arbitrary observer. We expressed the postulate in the following way:

PI. *Any equivalent observer M is furnished with a three-dimensional 'space-frame' S with origin M and a one-dimensional 'time-frame' T , and*

can give one and only one set of space coordinates (x, y, z) and time coordinate (t) to any point event E to within frame transformation.

Let me first explain what is meant by frame transformation. We regard two observers relatively at rest as essentially identical. And we call frame transformation the transformation between the frames of identical observers, that is, the frames relatively at rest to each other as well as such transformations of the time axis that simply change the scale of the time frame, namely, regraduation.

The postulate requires that an observer can give to an event a set of four real numbers representing coordinates (x, y, z, t) which is uniquely determined to within frame transformation.

It follows that there exists a relation between the coordinates (x, y, z, t) given to an event by an observer and the coordinates (x', y', z', t') given to the same event by another equivalent observer in his own frame. The relation is

$$(1) \quad x'^i = f^i(x^j), \quad |\partial x'^i / \partial x^j| \neq 0, \quad (i, j = 1, 2, 3, 4), \\ (x^1, x^2, x^3, x^4) \equiv (x, y, z, t).$$

In the above PI, we assumed the existence of a three-dimensional space-frame and a one-dimensional time-frame. However, it is not necessarily required that the two frames be combined to form a four-dimensional space-time. In this sense, this postulate may not be relativistic. Therefore, the postulate can cover both relativistic and non-relativistic theories. Namely, the postulate is not characteristic of relativistic theories. In fact, there are some transformation groups for which we can find no four-dimensional space-times satisfying the postulate of equivalency.

The second postulate we make requires that any observer can observe another observer. PI permits an observer to assign a set of coordinates to any point event, but it does not necessarily follow from this that the observer can do the same to another observer. The second postulate is necessary for this reason. It is the following.

PII. Any observer M can observe all other equivalent observers and they are all in motion relative to M .

Questions may arise as to what is meant by being in motion. Here as in the ordinary case, we say an observer is in motion relative to M if the spatial coordinates of the observer, (x, y, z) , are changing with time t .

The third postulate is a very important one.

PIII. *The group of frame transformations \mathfrak{G}_0 is given by the rotations*

$$R_1 = z\partial_y - y\partial_z, \quad R_2 = x\partial_z - z\partial_x, \quad R_3 = y\partial_x - x\partial_y$$

and the translations

$$T_1 = \partial_x, \quad T_2 = \partial_y, \quad T_3 = \partial_z$$

of the space frame S_M of M and the translation $U = \partial_t$ of time frame T_M of M . And this \mathfrak{G}_0 together with the set of transformations given by (1) forms a continuous group of transformations \mathfrak{G} .

The first question concerning this postulate will be why this particular transformation was chosen as the frame transformation. In our work we use coordinates without attaching any special meaning to them. Mathematically that should be satisfactory. However, we must examine the physical meaning of coordinates in order to compare the theory with the actual world in some way, or to apply the theory to observations of phenomena.

If the above mentioned frame transformation R_i , T_i , U can be interpreted as expressing the isotropy and homogeneity of space and the stationary character of time, then quite naturally, we can regard (x, y, z) as the cartesian coordinates of the euclidean space and t as the coordinate of time flowing uniformly. Certainly three-dimensional Riemannian space whose fundamental tensors are form invariant under coordinate transformations R_i and T_i is euclidean. This can be easily confirmed. In our work we have not assumed the metrical structure of space and time. Here we shall, however, postulate tentatively that the physical world forms four dimensional space-time. There may exist several ways to determine the structure of this space-time. Here we shall take, as an example, the following one tentatively.

Let us first notice that the following postulate we shall take here completely determines the structure of space-time in which equivalent observers can exist, and also the scale and the clock of that space-time. Namely, we postulate that the metric ds of the space-time be form invariant under the group \mathfrak{G} which is composed of the frame transformation \mathfrak{G}_0 and the transformation among equivalent observers as given by eq. (1). That is to say, we require that the space-time has the metric ds^2 given by

$$ds^2 = g_{ij}dx^i dx^j \quad (i, j = 1, 2, 3, 4)$$

with g_{ij} which is form invariant under \mathfrak{G} . Then, the laws of nature, if they

can be expressed as tensor equations, will be form invariant under \mathcal{G} . Thus, the laws of nature will assume the same expression for equivalent observers. This is the actual meaning of the equivalent observers.

We should also notice the following. Namely as will be shown later by an example, we found that for certain \mathcal{G} 's, there exists no four-dimensional space-time of the nature mentioned just now. In such cases, any two observers connected by \mathcal{G} in any four dimensional space-time whatsoever, will not be equivalent in the above sense. In such cases, we may take a viewpoint different from that of usual relativistic theories and say that there exists no four-dimensional space-time. How to interpret such an extraordinary case must be determined in each case.

Now let us return to the main story. The next postulate is:

PIV. *If M and M' are any two equivalent observers, they are in radial motion with respect to each other, and, furthermore, if M observes any E on the straight line MM' , then M' also observes the same E on the straight line $M'M$, independently of each time coordinate t and t' . Here, a straight line means the set of all the points invariant under any rotation of S .*

Implicit in this postulate is an assumption that we can treat three-dimensional space in analogy with one-dimensional space. Certainly this assumption will be natural. However, there are things characteristic of one-dimensional space. Therefore, we need to be careful.

Here I shall only mention the results obtained from the postulates I discussed so far, and shall not explain the actual calculations we did.

We found that the transformations between equivalent observers thus obtained were classified into the following three types. They are:

(a) Lorentz-type transformation

$$x' = (x - vt)/\sqrt{1 - av^2}, \quad y' = y, \quad z' = z, \quad t' = (t - avx)/\sqrt{1 - av^2}.$$

(b) Galilei transformation.

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t.$$

(c) K -transformation (as named by Takeno).

$$x' = x - v \exp(at), \quad y' = y, \quad z' = z, \quad t' = t.$$

It is very interesting that we obtained Lorentz-type transformation without any assumption on relative motion of observers. I will discuss this point later. Here I shall discuss the K -transformation. A characteristic feature of this transformation is that a point at rest in system S' moves

in (ST) system with the velocity proportional to the distance between the two origins of S and S' systems. Namely, we obtain from the above equation

$$[dx/dt]_{x'=\text{const.}} = a[x]_{x'=0}.$$

This relation reminds us of the velocity distance relation of nebular motion. If we choose a as Hubble's constant, this expression can be interpreted as the Hubble's relation in steady-state theory due to Bondi and Gold [5]. Assuming that we regard the postulate PIII as expressing the isotropy and homogeneity of space as well as the uniformity of time, it may be interesting to consider the relation between the assumption of invariance of the laws of nature for K -transformation and the perfect cosmological principle in the steady-state theory of cosmology. Thus we may say that PIII satisfies in a sense the conditions required by the perfect cosmological principle. In other words, we may say that PIII expresses the essential content of the perfect cosmological principle. Furthermore, there are many questions concerning the K -transformation like: what invariant relations do we have under this transformation? or what kind of dynamics corresponds to this transformation? We are now studying the applicability of the transformation to cosmology.

Lastly we shall remark on some problems concerning the structure of space and time. An especially remarkable feature of the K -transformation is that there exists no four-dimensional space-time of which the metric is form invariant under the group \mathcal{G} comprising the K -transformation. Hence, it is not proper to imagine in the above stated sense a four-dimensional space-time as the background in which we consider equivalent observers connected by the K -transformation. We, therefore, expect that a cosmology completely different from the relativistic one will come out if we adopt this transformation.

3. Equivalent Observers in Special Relativity. Now I want to change my subject to the work I did on equivalent observers in special relativity. The main problem is how to axiomatize fundamental principles of special relativity. Let us consider first the special principle of relativity. How to express this principle differs somewhat from person to person. Here, I borrow from the statement by Einstein himself [6].

If K is an inertial system, then every other system K' , which moves uniformly and without rotation to K , is also an inertial system: the laws of nature are in concordance for all inertial systems.

The principal concepts which should be examined in this principle are the following: first, inertial system and uniform motion; then what is actually meant by the statement that the laws of nature are in concordance for all inertial systems. In my paper [4], I discussed mainly this principle and did not touch the principle of constancy of light velocity.

Now we shall try to axiomatize the special principle of relativity. At the beginning we postulate the existence of observers, space-frame and time-frame. First, we make the same postulate as PI we gave before in Section 2. We shall call it AI here.

AI. Any equivalent observer M is furnished with a three-dimensional 'space-frame' S with origin M and a one-dimensional 'time-frame' T , and can give one and only one set of space coordinate (x, y, z) and time coordinate (t) to any point event E to within frame transformation.

By this postulate, it becomes possible to correspond a set of space coordinate (x, y, z) and time coordinate (t) to any point event. The postulate specifies three dimensionality of space and one dimensionality of time. An important conclusion of relativity tells that the space and time cannot be separated as two independent objective entities. However, it does not follow from this conclusion that the space and time cannot be separated for each individual observer. Hence, our postulate is not in contradiction with the existence of the space-time in relativistic sense. From AI we can conclude that there exist different observers and coordinate transformation between their space and time frames.

Next we adopt PII stated in Section 2 and call it AII.

AII. Any observer M can observe all other equivalent observers and they are moving relative to M .

Thirdly, we postulate the existence of uniform motion. This is the central point of the theory.

AIII. There exist point events which move uniformly.

Instead of postulating the existence of uniform motions as done here, we could have postulated the existence of clock and scale to define the structure of space and time, and could have obtained the same result. However, we want to use only kinematical concepts at the beginning. Now questions arise as to what objects make uniform motion and also as to how one can recognize uniform motion. The answer to these could be given by introducing dynamical concepts. For instance, one could define

uniform motion from the absence of external forces. However, if we want to proceed following this line of thought, dynamical aspects must be postulated first. Here we shall not, however, do this. The actual problem here will be how to express the uniform motion in the space-and-time-frame. Next we shall consider this problem.

By AI each observer was given a space frame and a time frame. However, there still remained the degree of freedom of the frame transformations. Using this freedom, we shall choose the space and time frames so that we can express uniform motion in a simple way.

DEFINITION 1. *We call a coordinate system a NORMAL FRAME if the coordinates (x, y, z, t) of a point event in uniform motion satisfy the following relations in this frame.*

$$(2) \quad x = v_x t + c_x, \quad y = v_y t + c_y, \quad z = v_z t + c_z.$$

Here v 's and c 's are constants. By these relations, we have now an expression for uniform motion. Now we shall consider the frames which are in uniform motion. In the following, we shall exclusively deal with normal frames.

DEFINITION 2. *If any point at rest in frames $(S'T')$ of an observer M' has always the coordinates that satisfy the relation (2) with the same v 's in frame (ST) of another observer M , then frame $(S'T')$ is IN UNIFORM MOTION RELATIVE TO (ST) .*

The existence of such a normal frame can be a question. That is to say, we are given the uniform motion by postulate, but it is not guaranteed that we can always find a frame in which we can express the uniform motion by equation (2). Hence, we shall assume the existence of a normal frame.

AIV. *To each observer, there exists a normal frame.*

The next axiom is a keypoint of the special principle of relativity.

AV. *Any normal frame which can be obtained by frame transformation from a normal frame (ST) or any normal frame which is moving uniformly relative to (ST) is equivalent to (ST) .*

The word "equivalent" used in the above AV means that the laws of nature are in concordance for the frames under consideration. We shall postulate the following set of axioms for equivalency. These hold for the

usual equality relation. Writing $A \equiv B$ to express that A is equivalent to B , we shall postulate the following relations:

AVI. *Axiom of equivalence.*

- (i) $A \equiv A$,
- (ii) if $A \equiv B$, then $B \equiv A$,
- (iii) if $A \equiv B$ and $B \equiv C$, then $A \equiv C$.

From the above AVI we can easily derive the following theorem.

THEOREM I. *Coordinate transformations between equivalent frames form a group.*

From the above axioms we can obtain the explicit form of the coordinate transformation from one normal frame to another. It is

$$(3) \quad x'^i = a_j^i x^j + c^i, \quad \det(a_j^i) \neq 0, \quad (i, j = 1, 2, 3, 4).$$

Here a 's and c 's are constants. As is well known, these transformations form the affine group. Therefore we obtain the following theorem.

THEOREM 2. *The set of transformations between normal frames forms the affine group.*

Evidently, this group includes as a sub-group the group of frame transformations.

If we further want to derive the constancy of the velocity of light, we have to define clock, scale or the metrical structure of space and time. By suitable stipulation of these concepts, we shall obtain the Lorentz transformations.

Before proceeding further, I want to come back to the problem of how to define uniform motion. The linear form we adopt was of course in direct analogy with euclidean space. Of course, there is no a priori reason for euclidean space. However, that the euclidean space is plausible may be seen as follows. In order to discuss the structure of space and time, we will have to introduce the metric of the space. Let us assume that the metric dl of (x, y, z) space is given by

$$dl^2 = g_{ij} dx^i dx^j, \quad (i, j = 1, 2, 3), \quad (x^1, x^2, x^3) = (x, y, z).$$

It will be quite natural to assume that the distance dl , which a point in uniform motion travels in time dt , is proportional to dt . If we assume this, then g_{ij} must be constant. From this we can easily prove the euclidean

property of the space. Then, we can introduce a cartesian coordinate system, and can define scale. Clock can be defined by combining scale and uniform motion.

In pre-relativistic theories, it is postulated that the running rate of a clock is the same for all observers, independently of their state of motion. Namely, the existence of an absolute time lapsing objectively is assumed. We do not make such an assumption, since there is no compelling reason for this. The running rate of the moving clock can be determined by (3), namely by its state of motion and the nature of scale which is determined by the euclidean nature of the space.

The axiomatic formulation of the special principle of relativity has been the main problem of the foregoing discussions. Our papers were attempts aimed at this end. Of course, we did not aim at rigorous axiomatization of the theory. Our interest was not in logical exactness but was rather in knowing how to express the content of the special principle of relativity. We believe that any attempt to axiomatize special relativity should start from analyzing the content of the special principle of relativity in all possible ways.

Our work reveals that uniform motion, normal frame and Minkowski space-time are cyclically related and that logically there is no reason to give priority to one of them. Therefore, either to assume the existence of objects which undergo uniform motion first, or to assume Minkowski space-time first, will be a kind of tautology.

If we want simplicity and rigor in the axiomatization of special relativity, then the existence of Minkowski space-time will have to be postulated first. Or to postulate the constancy of light velocity first instead of doing it last may be a simpler way than to specify the nature of space-time first. Whichever way we choose, there remains a number of problems to be considered in axiomatization of special relativity. Our work will serve to solve one of these problems; however, our work has the following weak point. Namely, the weakest point of our paper lies in not drawing any conclusion about how to specify the space-time structure. On the other hand, because of this deficiency we are left with the freedom of choosing a space-time structure. This is the next problem to be studied.

Bibliography

- [1] GÖDEL, K., *An example of a new type of cosmological solutions of Einstein's field equations of gravitation*. Reviews of Modern Physics, vol. 21 (1949), 447–450.
——, *A remark about the relationship between relativity and idealistic philosophy*, in SCHILPP, P. A. (ed.) *Albert Einstein: Philosopher-Scientist*. New York 1951,— pp. 555–562.
GRÜNBAUM, A., *Das Zeitproblem*. Archiv für Philosophie, vol. 7 (1957), pp. 165–208.
- [2] NARIAI, H., *On a new cosmological solution of Einstein's field equations of gravitation*. The Science Reports of the Tôhoku University, Ser. I, vol. XXXV (1951), pp. 62–67.
- [3] UENO, Y. and H. TAKENO, *On equivalent observers*. Progress of Theoretical Physics, vol. 8 (1952), pp. 291–301.
- [4] ———, *On the equivalency for observers in the special theory of relativity*. Progress of Theoretical Physics, vol. 9 (1953), pp. 74–84.
- [5] BONDI, H., *Cosmology*. Cambridge 1952, 146 pp.
- [6] EINSTEIN, A., *The meaning of relativity*. Princeton 1953, 25 pp.

ON THE FOUNDATIONS OF QUANTUM MECHANICS ¹

HERMAN RUBIN

University of Oregon, Eugene, Oregon, U.S.A.

1. We shall consider several formulations of the foundations of quantum mechanics, and some of the mathematical problems arising from them. Various of these problems will be treated in greater or less detail.

Most of the results presented here are not new, and it is the purpose of this paper mainly to bring to the attention of the worker in this field some of the difficulties which they have blithely overlooked. Most of the mathematicians dealing with the foundations of quantum mechanics have concerned themselves mainly with Hilbert space problems; one point they have brought out is the distinction between pure and mixed states. We shall not concern ourselves here with this problem, but shall confine our attention to pure states.

We give three formulations in detail; A, the Hilbert space formulation with unitary transition operators, B, the matrix-transition-probability-amplitude formulation, and C, the phase-space formulation. Each of these formulations is adequate for quantum mechanics. In formulation A in the classical case, the problem is usually specified by specification of the Hamiltonian and then solved by means of the Schrödinger equation; Feynman has proposed a method of path integrals which are not, as claimed, the average over a stochastic process, and, while a similarity to stochastic processes exists and should be exploited, does not mean that theorems and methods applicable in stochastic processes automatically apply. The same remarks apply to approach B, and a table is included of some important differences between stochastic and quantum processes. The identifiability problem is also pointed out for formulation B.

Formulation C is formally much closer to stochastic processes than A or B, but important differences are apparent. First and most important, the joint "density" of position and momentum need not be non-negative or even integrable. This, it seems to the author, implies that not only are position and momentum not simultaneously precisely measurable, but

¹ Research partially supported by an OOR contract. Reproduction in whole or in part is permitted for any purpose of the United States government.

that they are not even simultaneously measurable at all. It is true that non-negativeness of the density is preserved, but even here the motion is not that of a stochastic process.

2. Let \mathcal{H} be a Hilbert space, \mathcal{S} a partially ordered set — which in the relativistic case could be thought of as the set of all space-like surfaces, and in the classical case all points of time. Suitable conditions which will not be discussed here are to be imposed on \mathcal{S} .

A. For all $S, T \in \mathcal{S}$, $S < T$, there is a unitary operator U_{TS} on \mathcal{H} such that if $R < S < T$,

$$(1) \quad U_{TR} = U_{TS}U_{SR}.$$

In the classical case

$$(2) \quad U_{TS} = \exp\left(\frac{-i}{\hbar}(T - S)\mathbf{H}\right),$$

where \mathbf{H} is the Hamiltonian, and the Hilbert space may be taken to be L_2 over a Euclidean space of suitable dimensionality.

A central problem in quantum mechanics is specification of the Hilbert space and unitary operators involved.

Let E and F be complete spectral decompositions of the identity. Since for all $x \in \mathcal{H}$, $x = \int dEx = \int dFx$, we have $U_{TS}x = \iint dF U_{TS}dEx$, integrated first over E . But this is just the formulation of matrix mechanics. Thus if suitable regularity conditions are satisfied,

B. For all $R, S, T \in \mathcal{S}$, $R < S < T$, D, E, F complete spectral decompositions of the identity,

$$(3) \quad dF_T = \int d\lambda_{TS}(F, E)dE_S,$$

and

$$(4) \quad d\lambda_{TR}(F, D) = \int d\lambda_{TS}(F, E)d\lambda_{SR}(E, D).$$

One can reconstruct U from λ .

If the spectral decompositions are discrete, the integration becomes a summation. Also, we have the following interpretation of λ : the probability that an observation at "time" T will yield a result in a set \mathcal{X} given that an observation at "time" S yields a result E is

$$(5) \quad \int_{\mathcal{X}} |d\lambda_{TS}(F, E)|^2.$$

This has been interpreted as analogous to a stochastic process. However,

the differences are quite apparent to one familiar with stochastic processes, and are important. For a stochastic process, the analogues of (3) and (4) are customarily taken as definitions. However, expression (5) is replaced by

$$(6) \quad \int_{\mathcal{X}} \lambda_{TS}(F, E) dF.$$

The analogue of approach A is not as immediate. \mathcal{H} is to be replaced by an L_1 space over a finite measure space, which can be abstractly characterized. Then U_{TS} becomes a positive linear operator on \mathcal{H} to \mathcal{H} and (1) is satisfied. In addition, for some strictly positive function f_1 , and all S and T , $U_{TS}f_1 = f_1$. Also we may frequently, but not always, in the stationary classical case, write

$$(7) \quad U_{TS} = \exp[(T - S)V],$$

where V is called the infinitesimal generator of the semigroup U .

To see the differences clearly, let us consider the classical case where the Hilbert space is l_2 , i.e., all sequences of real numbers with finite sums of squares. Complex Hilbert space seems natural in quantum mechanics, but since every Hilbert space is automatically a real Hilbert space, and the analogy is better, we could use the real case. However, the complex case actually provides a closer analogy to a real stochastic process! If we now take $E = F$ to be the natural decomposition of l_2 , we may make the following analogy with discrete-space stochastic process. Starred sections refer only to stationary processes with linear "time".

Stochastic process	Quantum mechanics
Markov matrix U_{TS}	Unitary matrix U_{TS}
Transition probability u_{TSij}	Transition probability $ u_{TSij} ^2$
*Infinitesimal generator does not always exist and is not always unique.	*Infinitesimal generator always exists and is unique.
*In the regular case, the infinitesimal generator has all row sums 0, and all nondiagonal elements non-negative.	*Infinitesimal generator is a skew Hermitian matrix.
Ordering of \mathcal{S} irreversible.	Ordering of \mathcal{S} reversible.
*Trivial if periodic.	*Can be non-trivial and periodic.

From A , if the Hilbert space is explicitly an L_2 space, it may be

possible to write for a dense set of functions

$$(8) \quad U_{TS}(x) = \int K_{TS}(u, v) x(v) dv,$$

where K_{TS} is a unitary kernel. It may be possible, and indeed in the classical case it is, to determine the T -derivative of K at $T = S$. Suppose K_{TS}^* is a unitary approximation to K_{TS} , such that the T -derivatives of K and K^* coincide at $T = S$. In the classical case, Feynman did this by writing

$$(9) \quad K_{TS}^*(u, v) = N(T - S) \exp\left(\frac{i}{\hbar} A_{TS}(u, v)\right),$$

where $A_{TS}(u, v)$ is the action along the classical path from v at "time" S to u at "time" T . Then we may define U_{TS}^* from K_{TS}^* in a manner analogous to (8). It may be that

$$U_{TS} = \lim_{n \rightarrow \infty} \prod_{i=1}^n U_{T_i, T_{i-1}}^*, \quad T_0 = S, \quad T_n = T, \quad T_{i-1} \leq T_i,$$

when the partition becomes fine. Although there are several treatments in the literature, including some by prominent mathematicians, the existence and value of this limit has not been proved. From the Schrödinger equation, one can prove the following

THEOREM: *If there exists a basis of L_2 such that for each function x in the basis, the second derivatives of $U_{TS}x$ has a uniformly integrable Fourier transform, then $U_{TS} = \lim_{n \rightarrow \infty} \prod_{i=1}^n U_{T_i, T_{i-1}}^*$ where $T_0 = S, T_n = T, T_{i-1} \leq T_i$, and the partition becomes fine.*

It seems likely that this result can be considerably extended.

If we examine the analytic form of (9), we find that it resembles that of a diffusion process. However, the "variance" of the "diffusion process" would have to be purely imaginary. Furthermore, there are even periodic models in quantum mechanics which satisfy the theorem above. If $T - S$ is a multiple of the period, K_{TS} cannot be a function in the ordinary sense. In fact, if $T - S$ is a multiple of any discrete spectral value, this difficulty arises.

Another difficulty with this formulation is the statement that in the limit K_{TS} is the normalized mean value on x of $\exp(A(u, v, x))$ where x is a path with end points v at S and u at T . In the case of a diffusion process, it is well known that the corresponding exponent is infinite with proba-

bility one. The same difficulty has already been noted in the quantum-mechanical formulation.

The computation of the Feynman expression also is rather difficult to evaluate. However, stochastic process methods may be useful. While the process has purely imaginary variance, we may compute the diffusion process with real variance and use analytic continuation. Again, it remains to be proved that this method is correct. An intermediate approach would be to apply analytic continuation to the coefficient of the kinetic energy term alone. This last method has worked for the free particle and the harmonic oscillator, and methods for computing the results in general have been given by Kac.

One merit of the Feynman approach is that it has great possibility of generalization in that it leads to a specific result for U_{TS} , the specification of which is a main problem of quantum mechanics and usually overlooked by mathematicians dealing with the subject.

There is an outstanding question which arises from the empirical standpoint; namely, if the model is correct, how much of the model can be determined by even an infinite number of observations? This seems to be most clearly brought out in formulation B above. For simplicity, let us assume that the decompositions E and F are discrete. Then the observable quantities are $|\lambda_{TSij}|^2$. Clearly these are not always adequate for fixed E and F even if S and T are arbitrary.

In the discrete case, $\lambda_{TSij} = (U_{TSfi}, e_j)$. If we may vary E arbitrarily, we may determine U_{TSfi} completely apart from a constant of absolute value 1 for each i . If furthermore $E = F$ and for almost all $S, T, U_{TSij} \neq 0$ for all i and j , we can determine U_{TSij} apart from a constant of absolute value l independent of i and j , i.e., apart from a gauge transformation.

Another approach is the statistical approach of Moyal. This approach, originally due to Wigner, is to investigate the joint "distribution" of position and momentum. First, suppose a finite number A_1, \dots, A_n of Hermitian operators are given. Then if they have a joint distribution, its characteristic function is $E(\exp \sum i t_j A_j)$. However, the operator inside the expectation is a unitary operator, and consequently the expectation in question exists.

Therefore we should be able to determine the distribution from the expectation. For example, let A_1, A_2 , and A_3 be the spin operators for an electron in a hydrogen atom about which nothing has been deduced by experimentation about the spin. Then $E(\exp \sum i t_j A_j) = \cos \frac{\hbar}{2} \sqrt{\sum t_j^2}$,

which is certainly not the characteristic function of any distribution. Let us proceed as if this difficulty does not arise, and let us treat the case of position and momentum. We obtain the characteristic function

$$(10) \quad E(\exp(i\alpha p + i\beta q)) = \int \psi^*(q - \tfrac{1}{2}\hbar\alpha) e^{i\beta q} \psi(q + \tfrac{1}{2}\hbar\alpha) d q,$$

and the corresponding density

$$(11) \quad f(p, q) = \frac{1}{2\pi} \int \psi^*(q - \tfrac{1}{2}\hbar\alpha) e^{-i\alpha p} \psi(q + \tfrac{1}{2}\hbar\alpha) d\alpha.$$

Another example of the misbehavior of f is in order. Let us consider a plane wave passing through a slit of aperture $2a$. Then $\psi(x) = \frac{1}{\sqrt{2a}}$, $-a \leq x \leq a$, and we obtain

$$(12) \quad f(p, q) = \begin{cases} \frac{1}{2\pi a p} \sin \frac{2(a - |q|)p}{\hbar} & |q| \leq a, \\ 0 & |q| > a. \end{cases}$$

We clearly see that f is not non-negative, and not even Lebesgue integrable.

It would be desirable to have an abstract characterization of all permissible "densities", as the density is adequate both for the kinematics and for the dynamics of quantum mechanics. Let us proceed to do so. As to the kinematics, it follows from (11) that for almost all x, y ,

$$(13) \quad \psi(x)\psi^*(y) = \int f\left(p, \frac{x+y}{2}\right) e^{i(x-y)p/\hbar} dp.$$

Therefore

$$(14) \quad \begin{aligned} \int \int f\left(p, \frac{x+y}{2}\right) f\left(\pi, \frac{z+w}{2}\right) e^{i[(x-y)p + (z-w)\pi]/\hbar} dp d\pi \\ = \int \int f\left(p, \frac{x+w}{2}\right) f\left(\pi, \frac{z+y}{2}\right) e^{i[(x-w)p + (z-y)\pi]/\hbar} dp d\pi \end{aligned}$$

for almost all x, y, z, w . If, in addition, $f(p, x)dp$ is a probability density, there will be a unique solution for ψ apart from a factor of absolute value 1. Conversely, if f satisfies (14) and $f(p, x)dp$ is a probability density, the ψ deduced from f by (13) yields f in return.

Concerning the dynamics of the process, Moyal has shown that the

temporal derivative of the characteristic function (10) is, where H is the classical Hamiltonian,

$$(15) \quad \frac{1}{\hbar} \int \int [H(p + \tfrac{1}{2}\hbar\beta, q - \tfrac{1}{2}\hbar\alpha) - H(p - \tfrac{1}{2}\hbar\beta, q + \tfrac{1}{2}\hbar\alpha)] f(p, q) e^{i(\alpha p + \beta q)} d p d q.$$

Inverting this, we obtain for the derivative of the density

$$(16) \quad \frac{\partial f(p, q, t)}{\partial t} = \frac{1}{\pi^2 \hbar^3} \int \int \mathcal{J}(e^{2i(\mathbf{p}\mathbf{q}' - \mathbf{q}\mathbf{p}')/\hbar} \hat{H} \left(-\frac{2}{\hbar} (q - q'), \frac{2}{\hbar} (p - p') \right) f(p', q', t) d p' d q',$$

where $\mathcal{J}(u + iv) = v$, and \hat{H} denotes the Fourier transform of H . A more convenient form of (16) is

$$(17) \quad \frac{\partial f(p, q, t)}{\partial t} = \frac{1}{4\pi^2 \hbar} \int \int \mathcal{J}(e^{i(\mathbf{p}\beta + \mathbf{q}\alpha)} \hat{H}(\beta, \alpha)) f(p - \tfrac{1}{2}\hbar\alpha, q + \tfrac{1}{2}\hbar\beta) d\alpha d\beta.$$

Even this form gives some difficulties in evaluation because of the non-existence in the usual sense of \hat{H} , and the right-hand side of (16) has to be evaluated by approximations. The form which Moyal seems to prefer is even worse in this respect, but it also has some advantages.

$$(18) \quad \frac{\partial f(p, q, t)}{\partial t} = \frac{2}{\hbar} \sin \frac{\hbar}{2} \left[\frac{\partial}{\partial p_f} \frac{\partial}{\partial q_H} - \frac{\partial}{\partial p_H} \frac{\partial}{\partial q_f} \right] H(p, q) f(p, q, t).$$

[This latter form shows more clearly the relationship between classical and quantum mechanics, but the differential operator on the right is of infinite order and analytic difficulties may clearly ensue. In the case in which H is a polynomial of degree at most 2, (18) reduces to the classical equations of motion; quantum-mechanical considerations come in only through restrictions (14) on f .]

In any case, it follows that the dynamics of the phase-space representation above does not further involve the wave function. Consequently, the dynamics of ψ is determined up to a gauge transformation by equation (17), and hence the following formulation is adequate for classical one-dimensional quantum mechanics:

C. *There is a function f of three arguments satisfying almost everywhere for some value t of its third argument, (14) and $\int f(p, x, t) dp$ is a probability density, and satisfying (17).*

It is clear how to extend this to higher dimensional cases.

This "probabilistic" procedure might also be used to construct the unitary kernel K_{ST} for the Feynman approach, although this has not been done.

Bibliography

- [1] FEYNMAN, R. P., *Space-time approach to nonrelativistic quantum mechanics*. Review of Modern Physics, vol. 20 (1948), p. 367.
- [2] GELFAND, I. M. and A. M. YAGLOM. *Integration in function spaces and its application to quantum physics*. Uspekhi Matematicheskikh Nauk (N.S.), vol. 11 (1956), p. 77.
- [3] KAC, M., *On some connections between probability theory and differential and integral equations*. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley 1951.
- [4] MONTROLL, E. W., *Markoff chains, Wiener integrals, and quantum theory*. Communications on Pure and Applied Mathematics, vol. 5 (1952), p. 415.
- [5] MORETTE, C., *On the definition and approximation of Feynman's path integrals*. Physical Review, vol. 81 (1951), p. 848.
- [6] MOYAL, J. E., *Quantum mechanics as a statistical theory*. Proceedings of the Cambridge Philosophical Society, vol. 45 (1949), p. 99.
- [7] SEGAL, I. E., *Postulates for general quantum mechanics*. Annals of Mathematics (2), vol. 48 (1947), p. 930.
- [8] STONE, M. H., *Notes on integration I, II, III, IV*. Proceedings of the National Academy of Sciences, U.S.A., vol. 34 (1948), p. 336, p. 447, p. 483, vol. 35 (1949), p. 50.

THE MATHEMATICAL MEANING OF OPERATIONALISM IN QUANTUM MECHANICS

I. E. SEGAL

University of Chicago, Chicago, Illinois, U.S.A.

1. Introduction. An operational treatment may be described as one that deals exclusively with observables; but the latter term is physically as well as mathematically somewhat ambiguous. Our aim here is to circumscribe this ambiguity by axioms for the observables that will be satisfactory as far as they go, but by no means categorical. On the other hand, it will turn out that it is not too far from such axioms to plans for a categorical model representing the field of all elementary particles.

The need to consider so broad a system arises in several ways. For one thing, no axiom system is secure if it does not treat a closed system, and except substantially in the case of classical quantum mechanics (by which we mean the non-relativistic quantum mechanics of a finite number of degrees of freedom), there is no mathematical or physical assurance that the systems conventionally considered are really closed. In fact the evidence, — highly inconclusive as it may be, — points very much in the other direction. For another, although the mathematical foundations of classical quantum mechanics are in a relatively satisfactory state from at least a technical point of view (the theory is consistent, within obvious limits categorical, and realistic), time and energy play crucial but puzzling roles, as observables unlike the others. While this remains true in relativistic quantum field theory, for different reasons, it seems fair to say that one of the accepted informal axioms of the theory is that it must ultimately contain the solution to the puzzle, if such exists.

We should not gloss over the question of just what is a quantum field theory, — in fact, this is the main question we wish to examine here. It is a difficult question, since at present what we have, after thirty years of intensive effort, is a collection of partially heuristic technical developments in search of a theory; but it is a natural one to examine axiomatically. Present practice is largely implicitly axiomatic, and nothing

resembling a mathematically viable explicit constructive approach has yet been developed. In any event a constructive approach must presumably describe the physical particles with which an operational theory must deal in terms of the only remotely operational bare particles, a problem that is relatively involved in the current non-rigorous treatments, and needs to be clarified by a suitable axiomatic formulation.

Description of a field, whether classical or quantum, involves analytically three elements: (a) its *phenomenology*, i.e. the statement of what mathematically are the observables of the field, and what are their physical interpretations, — including especially, in the case of quantum fields, the statistics, i.e. the observables called single-particle occupation numbers, which do not exist in classical fields, and form the basis for the particle interpretation of quantum fields; (b) its *kinematics*, i.e. the transformation properties of the field observables under the fundamental symmetry group of the system; (c) its *dynamics*, or the ‘temporal’ development of the field, where however the ‘dynamical time’ involved must be distinguished from the ‘kinematical time’ involved in (b). The dynamics results from the interaction between the particles constituting the field, and is in fact its only observable manifestation, while the kinematics has nothing to do with this interaction.

The present state of the axiomatics of these elements and of the desiderata relevant for further developments is discussed from a jointly mathematical and operational viewpoint in the following.

2. Phenomenology. This is the best-developed of the relevant phases of quantum mechanics from both a mathematical and an operational point of view. One knows that the bounded observables, which are the only ones that can in principle be measured directly, form a variety of algebra, of which the self-adjoint elements of a uniformly closed self-adjoint algebra of operators on a Hilbert space (C^* -algebra) is virtually the exclusive practical prototype. One knows also that the states of the system are represented by normalized positive linear functionals on the algebra, the value of such a functional on an element being what is conventionally called the ‘expectation value of the observable in the state’ in physics, but there being no operational distinction between the state and the associated functional, — i.e. operationally (and in our usage in the following) a state is precisely such a functional. In these terms the

essential notions of pure state, spectral value of an observable, probability distribution of an observable in a state, etc., can be axiomatized and shown to admit a mathematical development adequate for physical needs.

An important conclusion of the theory is that a physical system is completely specified operationally by giving the abstract algebra formed by the bounded observables of the system, i.e. the rules for forming linear combinations of and squaring observables. In particular, operationally isomorphic algebras of observables that are represented by concrete C^* -algebras on Hilbert spaces, do not at all need to be unitarily equivalent, even when, for example, they are both irreducible. The irrelevant and impractical requirement of unitary equivalence is in fact the origin of serious difficulties in the development of quantum field theory, a point with which we shall deal more explicitly later.

The subsumption of quantum fields under general phenomenology involves the formulation and treatment of the 'canonical field variables' and the 'occupation numbers'. Traditionally the former were an ordered set of symbols p_1, p_2, \dots and q_1, q_2, \dots satisfying the commutation relations that had been so successful in classical quantum mechanics. (This is for 'Bose-Einstein' fields; relevant also are 'Fermi-Dirac' fields, but as these involve no great essential novelty as far as the present aspects of axiomatics go, the present article treats only the Bose-Einstein case.) It was assumed that these were an irreducible set of self-adjoint operators, and that any two such systems were equivalent; upon this informal axiomatic basis the theory rested. But from the very beginning the success of quantum field theory was attended by 'infinities' in even the simplest cases, and more recently it has been found that there exist at least continuum many inequivalent irreducible systems of canonical variables. Such troubles made it uncertain whether the phenomenological structure described above was strictly applicable in the case of quantum fields, or at least whether the canonical variables really were self-adjoint operators in a Hilbert space. The proper sophistication, based on a mixture of operational and mathematical considerations, gives however a unique and transparent formulation within the framework of the phenomenology described; the canonical variables are fundamentally elements in an abstract algebra of observables, and it is only relative to a particular state of this algebra that they become operators in Hilbert space.

In a formal way it was easily seen that the symbolic operator $(p_k + iq_k)(p_k - iq_k)$ had integral proper values ($i^2 = -1$), and for this and related reasons could be interpreted as 'the number of particles in the field in the k th state', which is essentially what puts the 'quantum' into 'quantum field theory', by giving it a particle interpretation. Those particles, the 'quanta' of the field, have generally been presumed to be 'represented' by the vectors in a linear space, proportional vectors being identified. This linear space L does not have direct operational significance, since what is more-or-less directly observed are the 'occupation numbers of single-particle states', i.e. the observables just defined (formally). But the general principle that there exists (theoretically) a single-particle space L , spanned by an infinite set of vectors f_1, f_2, \dots , and such that $p_k + iq_k$ can represent in a certain sense the creation of a particle with 'wave function' e_k , and the operator defined above the total number of such particles in the field, has attained virtually as well-established a position as the general phenomenological principles described earlier. The great empirical success of relativistic quantum electrodynamics, in which the photon and the electron are represented by suitably normalizable solutions of Maxwell's and Dirac's equation, respectively, provides, among other developments a basis for this principle, and indicates also that L should admit a distinguished positive-definite Hermitian form, which determines, e.g., when two particles are empirically similar. It is conservative as well as useful in treating certain theories of recent origin to assume only a distinguished topological structure that may be induced by such a form, which turns out to involve no really significant weakening of the foundations, and ultimately to clarify their logical structure. In fact, partly for logico-mathematical reasons, and partly with a view to deriving ultimately the relevance of complex scalars for the single-particle space from invariance under so-called particle-anti-particle conjugation, it is appropriate to assume initially that the single-particle structure is given by an ordered pair of mutually dual, real-linear spaces with the topological structure described, and with which the canonical p 's and q 's are respectively associated. A distinguished admissible positive-definite inner product in one of these spaces will give a distinguished complex Hilbert space structure on the direct sum of the two spaces, but there are other ways in which this more conventional structure may arise.

Taking then a conservative position, and defining a phenomenological single-particle structure as an ordered pair of real-linear spaces (H, H')

that are mutually dual in the sense that there is given a distinguished non-singular bilinear form $x.y'(x \in H, y' \in H')$, a quantum field relative to this structure may be rigorously, but provisionally, described as an ordered pair of maps $(p(.), q(.))$ from H and H' respectively to the self-adjoint operators on a complex Hilbert space K , satisfying the 'Weyl relations':

$$\begin{aligned} e^{ip(x)}e^{iq(y)} &= e^{ip(x+y)}, & e^{iq(x')}e^{iq(y')} &= e^{iq(x'+y')} \\ e^{ip(x)}e^{iq(y')} &= e^{ix.y'}e^{iq(y')}e^{ip(x)}, \end{aligned}$$

which are formally equivalent to the conventional commutation relations, but mathematically more viable, in that difficulties associated with unbounded operators such as the p 's and q 's themselves, are avoided. This is merely an honest, if slightly sophisticated and general, mathematical transcription from the ideas and practice of physical field theory, but it is useful in providing a basis for deciding what is literally true about quantum fields, and what is figurative or symbolic. Thus the physical folk-theorem: 'Any two irreducible quantum fields are connected by a unitary transformation' is literally false, although it has figurative validity, which on the basis of a further mathematical development can be made rigorously explicit. The needs of field dynamics leads to this development and to a revision of the present provisional notion of quantum field which will be indicated later.

Also in need of revision is the definition of occupation number of a single-particle state. The validity of the occupation number interpretation of the given operator depends in part on the representation of the total field energy (etc.) in terms of occupation numbers of states of given energy, in keeping with the idea that it should equal the sum of the products of the various possible single-particle energies with the numbers of particles in the field having these energies. This holds for a certain mathematically and physically distinguished quantum field in the foregoing sense, studied by Fock and Cook, often called the 'free field', although actually of dubious application to free incoming physical fields, and almost certainly inapplicable to interacting fields. In any event, it breaks down in the case of arbitrary fields, and there has been some uncertainty as to whether a physically meaningful particle interpretation of an arbitrary field could be given. The solution to this problem depends on the proper integration of statistics with kinematics, to which we now turn.

3. **Kinematics.** It is axiomatic that a suitable displacement of the single-particle structure should effect a corresponding field displacement. In the case of a classical field, given say by Maxwell's equations, it is clear an arbitrary Lorentz transformation L induces a transformation $U(L)$ in the space of solutions. From a quantum-field-theoretic point of view however, $U(L)$ is merely a displacement in the single-particle space (of normalizable photon states), and what is needed is a transformation $V(L)$ on the field vector state space K of the preceding section. The assumption that $V(L)$ exists means essentially that any admissible change of frame in ordinary physical space should give a corresponding transformation on the field states. In addition, the assumed independence of transition probability rates of elementary particle processes from the local frame of reference has led to the further assumption that $V(L)$ is a projective unitary representation of the Lorentz group, in at least the case of the 'free incoming' physical field.

In addition to the Lorentz group, there is a group of transformations in the single-particle vector state space which plays an important part in nuclear physics, and which do not arise from transformation in ordinary physical space, — namely, transformations in isotopic spin space. In the absence of precise knowledge, it is assumed that this group acts independently of the Lorentz group, but its precise structure as an abstract group is undecided, and it is quite uncertain whether it is rigorously true that these transformations commute with the action of the Lorentz group on the single-particle space. There is also the group of gauge transformations, which is important in quantum electrodynamics, but does not have any counterpart in most other elementary particle interactions. The improper Lorentz transformations have recently been the subject of intense interest. These transformations give rise to outer automorphisms of the proper Lorentz group, and there seems to be at present no operational reason to doubt that this is their chief significance (rather than as direct transformations in ordinary space-time), but the experimental situation is far from giving any assurance that this is the case. In the case of standard relativistic theory, this leaves only charge and particle-anti-particle conjugation, of which the latter is connected with the equivalence between particle and the contragredient anti-particle transformations, and does not appear to represent in a natural way a group element. Finally, these and other kinematical loose ends, together with the dynamical divergences, have led certain scientists to investigate the

possibility that some other group may give more satisfactory results than the Lorentz group, just as this group gave ultimately a sounder theory than the Galilean group of Newtonian mechanics, and of which the Lorentz group will be a type of degenerate form, just as the Galilean group is a degenerate form of the Lorentz group.

On a conservative basis, it seems that about all that may legitimately be assumed of a mathematically definite character is that there exists a fundamental symmetry group G , which may reasonably be assumed to be topological, and which acts linearly and continuously on the single-particle vector state space. A priori it might appear that this is not sufficient as a basis for an effective field kinematics, but it turns out that special properties of G and of its action on the single-particle space are not significant as regards the foundations of field kinematics. The main desideratum is to establish the appropriate action of G on the field, and this exists substantially in all cases, provided it is the operational action that is considered. That is to say, the action of G on the state vectors of the field, — which in the case of standard relativistic theory is given formally in detail in the recent treatments of field theory in the literature, — does not need to exist in a mathematical sense, any more than it exists operationally; but the action of G on the field observables, which is formally to transform them by its action on the state vectors, has effective mathematical existence. However, to this end it is necessary to make the revision of the notion of quantum field referred to above, to which one is naturally led by dynamical and further operational consideration.

Before going into these matters, we mention that the generality of the foregoing approach to kinematics permits the integration of the statistics with the kinematics. Any non-singular continuous linear transformation on the single-particle structure (H, H') preserving the fundamental skew form $x.y' - u.v'$ (x and u arbitrary in H , y' and u' arbitrary in H') acts appropriately on the field observables; in particular certain phase transformations in the single-particle space so act, and the occupation numbers are obtained as generators of one-parameter groups of such field actions. A development of this type is needed for the particle interpretation of fields, if one is to avoid the ad hoc assumption that the free incoming physical field is mathematically representable by the special representation referred to earlier, as well as for dealing with the concept of bound state.

4. Dynamics. In conventional theoretical physics, a dynamical transformation is represented by a unitary transformation mathematically. In the case of an abstract algebra of observables as described above, it has however no meaning to say that a transformation of this algebra is given by a unitary transformation, for this may be true in certain concrete representations of the algebra and not in others. It is clear though that the transformation of the observables determined by a unitary operator in a concrete representation is an automorphism of the algebra. Since operationally an automorphism has all the relevant features of a dynamical (or, for that matter, kinematical) transformation, one is led to a generalization of conventional dynamics in which such a transformation is axiomatized as an automorphism of the algebra of observables. This is a proper generalization, in the sense that it is not always possible to represent an automorphism of an abstract C^* -algebra by a similarity transformation by a unitary operator in a given concrete representation space; but what is more relevant to field theory is that even when each of a set of automorphisms can be so represented, there will generally be no one representation in which all of the automorphisms are so representable.

This difficulty does not arise to any significant extent in the quantum mechanics of a finite number of degrees of freedom, for due to a special property of finite systems of canonical variables, every automorphism of the conventionally associated algebra of observables can, in any concrete representation, be induced by a unitary operator. But in the case of a quantum field, there are simple apparent dynamical transformations that can be shown to be not implementable by any unitary transformation in the case of the Fock-Cook field. Now there is no physical reason why every self-adjoint operator on the field vector state space should even in principle be measurable, but it has not been clear how to distinguish, in effective theoretical terms, those which were. To arrive at such a distinction, we consider that the canonical variables themselves should be measurable, and also, in accordance with conventional usage in the case of a finite number of degrees of freedom, any bounded 'function' of any finite set of canonical variables. However, since only finitely many particles are involved in real observations, other self-adjoint operators are only doubtfully measurable, except that uniform limits of such bounded functions must also be measurable, since their expectation value in any state is simply the limit of the expectation values of the approximating bounded functions. That is to say, uniform approximation is operationally meaning-

ful, since operators are close in this sense if the maximum spectral value of their difference is small. The point is now that the simple apparent dynamical transformations that could not be represented by unitary transformations in the field state space can however be represented by automorphisms of the algebra of observables just arrived at (e.g. division of the canonical p 's by $\lambda > 1$ and multiplication of the canonical q 's by λ can be represented by such an automorphism, although not by a unitary transformation in the Fock-Cook field).

More generally, the algebra of measurable field operators defined above is the same for all concrete quantum fields as defined earlier. That is, for any two quantum fields $(p(\cdot), q(\cdot))$ and $(p'(\cdot), q'(\cdot))$, relative to the same single-particle structure, there exists an isomorphism between the corresponding algebras that takes any (say, bounded Baire) function of $p(x)$ into the same function of $p'(x)$ for all x , and similarly for the q 's. This isomorphism is in fact unique, from which it can be deduced that any continuous linear single-particle transformation leaving invariant the fundamental skew form gives rise to a corresponding automorphism of the algebra. This resolves the problem of defining the field kinematics when the single-particle kinematics is given.

For an operational field dynamics we have to deal mainly (if not, indeed, exclusively) with the particular transformation that connects the so-called incoming and outgoing free fields, which may be defined as the scattering automorphism. In view of the uniqueness of the algebra of field observables, it does not matter in which representation this automorphism is given. Tied up with these notions are those of the physical vacuum state, physical particle canonical variables and occupation numbers, and the scattering operator. Since what is more-or-less directly observed for quantum field phenomena is interpretable as the scattering of an incoming field of particles, it is appropriate to attempt to formulate these various notions in terms of a given scattering automorphism ' s '. The physical vacuum state must certainly satisfy the condition of invariance under s . This will in general not give a unique state, but it is fairly reasonable to assume that in a realistic theory, the additional requirement of invariance under the kinematical action of a maximal abelian subgroup of the fundamental symmetry group may well give uniqueness. The axiom of covariance asserts that s commutes with the kinematical action of the entire symmetry group on the field observables, and from this and a well-known fixed-point theorem the existence of a physical vacuum as so defined follows.

Given a state of an abstract C^* -algebra that is invariant under an abelian group of automorphisms, there corresponds in a well-known mathematical manner, a concrete representation of the algebra on a complex Hilbert space K , and a unitary representation of the abelian group on the space, which give similarity transformations effecting the automorphisms. In this way there is determined the unitary scattering operator S , which in this particular representation implements the automorphism s , and a unitary representation of the maximal abelian subgroup of the covariance group. The vacuum state is represented by a vector of K , left invariant by S and this unitary representation. (In the application to standard relativistic theory, the abelian subgroup would consist of translations in space-time, which in conventional theory leaves only the physical vacuum fixed, among all physical states.) The incoming field is defined as that given by the representation, and the outgoing field as its transform under S , both having the vector state space K ; to avoid subtle and technical mathematical questions in this connection the physically plausible assumption of continuity of the physical vacuum expectation values of the $Ae^{itp(x)}B$ and $Ae^{itq(y')}B$, at least when x and y' range over finite-dimensional subspaces of the single-particle space, is made, where A and B are fixed but arbitrary field observables. The $\hat{p}(x)$ and $\hat{q}(y')$ that generate the homomorphic images of the one-parameter groups $[e^{itp(x)}: -\infty < t < \infty]$ and $[e^{itq(y')}: -\infty < t < \infty]$ are defined as the canonical variables of the free incoming physical field, and their transforms under S , those of the outfield. In defining single-particle state occupation numbers, it is convenient to assume present a distinguished complex Hilbert space structure in the direct sum $H + H'$. For any single-particle state vector x in $H + H'$, there is then a unique continuous one-parameter unitary group $[U(t): -\infty < t < \infty]$ taking x into $e^{it}x$ and leaving fixed the orthogonal complement of x . The corresponding automorphisms of the algebra of field observables likewise form a one-parameter group. In general they will not leave invariant the physical vacuum state, but again making physically plausible continuity and boundedness assumptions, there will be obtained finally a corresponding one-parameter group of linear transformations in K , which will have a 'diagonalizable' generator, i.e. one similar (in general, via a non-unitary transformation) to a self-adjoint operator. Although these occupation numbers are not self-adjoint, they have the crucial properties of having integral proper values; of being such that the total in-field energy, momentum, etc. the sum of the products of all single-particle energies,

momenta, etc. with the occupation numbers of the corresponding states in a formal, but partially rigorizable, manner; and of annihilating the physical vacuum state vector.

The fundamental problem of quantum field dynamics from an overall point of view is and always has been that of the so-called divergences. In present terms, this is the problem of establishing the existence of the scattering automorphism s , which must satisfy certain conditions, which however can not be stated with mathematical precision, this lack of precision being an inherent difficulty of the problem. That the present approach may well be relevant to this problem may be seen in the following way. The scattering automorphism may be given as an infinite product integral, and the crucial difficulty has always been that of establishing the existence of the integrand. This is given formally by a complex exponential of the integral at a particular time of the 'interaction Hamiltonian', whose character is relevant here only to the extent that in a variety of interesting and typical cases, it is a linear expression in the canonical p 's and q 's, whose coefficients are relatively untroublesome operators. E.g. in certain current theories of meson-nucleon interaction, they are simply finite-dimensional matrices; for fully quantized electrodynamics 'in a box' they are mutually commutative self-adjoint operators in a Hilbert space. Now there is no doubt that these formal operators are divergent, in the sense that they do not represent bona fide self-adjoint operators in the Fock-Cook representation, — in fact their domains in general appear to consist only of $\{0\}$. But in dealing with these formal operators, we are at liberty to change the representation employed for the canonical p 's and q 's according to the foregoing development. Now it can be shown that there always exists a representation for which $\Sigma_k q_k \times T_k$ represents in an obvious manner a bona fide hermitian operator, provided that each T_k is a bounded operator. One can deal similarly with $\Sigma_k (q_k \times T_k + p_k \times V_k)$ when the T_k and V_k are mutually commutative self-adjoint operators. In either case the complex exponential will be a well-defined unitary operator. Thus although the final physical results are independent of the representations employed in setting up the theory, the divergence or convergence, as operators in Hilbert space, of expressions involved in the analysis, may depend strongly on the representation.¹

¹ For a more detailed account of certain physical points, as well as references to proofs of relevant mathematical results, see Segal [2] and [3]. For another approach to the axiomatics of quantum field theory from a partially heuristic point of view, but with points of contact with the present approach, see R. Haag [1].

Bibliography

- [1] HAAG, R., *On quantum field theories*. Matematisk-Fysiske Meddelelser udgivet af det Kgl. Danske Videnskabernes Selskab. 29 (1955).
- [2] SEGAL, I. E., *The mathematical formulation of the measurable symbols of quantum field theory and its implications for the structure of free elementary particles*. To appear in the report of the International Conference on the Mathematical Problems of Quantum Field Theory (Lille, 1957).
- [3] —, *Foundations of the theory of dynamical systems of infinitely many degrees of freedom*. I. Matematisk-Fysiske Meddelelser udgivet af det Kgl. Danske Videnskabernes Selskab. 31 (1959).

QUANTUM THEORY FROM NON-QUANTAL POSTULATES

ALFRED LANDÉ

Ohio State University, Columbus, Ohio, U.S.A.

1. **Physical and Ideological Background.** Theoretical physics aims at deducing formal relations between observed data by the combination of simple and general empirical propositions which, if true, will 'explain' the variety of phenomena. In the process of constructing a physical theory on a postulational basis one may distinguish between three steps. *First*, by critical evaluation of experience one arrives at ideological pictures for the connection of individual data (e.g. for the 'path' of a firefly, Margenau) and at general notions expressed in everyday language which takes much for granted and may involve circularity in the definition of terms. *Second*, the resulting picture is formalized and condensed into general laws. *Third*, the formal laws are now put in correspondence with a physical 'model' which gives an operational definition of each symbol, resulting in a self-consistent physical theory. In spite of its vagueness, step 1 is of importance to the physicist since it furnishes a legitimate basis for his selection of one formalism among many possible ones as the formal substructure of his laws.

The quantum theory in its historical development has followed this procedure, its laws are based today on a few universal, though rather baffling, principles, the most prominent among them being those of *wave-particle duality*, *qp-uncertainty*, and *complementarity*. I submit, however, that the process of reduction has not gone far enough, and that the quantum principles just mentioned can be reduced further to simple empirical propositions of a *non-quantal* character, the combination of which yields the quantum principles as consequences. The latter can thus be 'explained' on an elementary and more or less familiar background "so that our curiosity will rest" (Percy Bridgman),. Conforming with step 1 above, I begin with considerations of a somewhat vague character in order to lay the ideological groundwork for the formal substructure of quantum mechanics. — Two objects, *A* and *B*, or two 'states' *A* and *B* of the same 'kind' of object, may be said to be different, written $A \neq B$, when *A* and *B* are discernible, i.e. separable by means of

some device, shortly denoted as a 'filter', responding to B with 'no' when $B \neq A$, and with 'yes' when $B = A$, as depicted by Figs. 1a and 1b where \bar{A} is written for different from A or *non-A*. The term 'state', 'filter', 'kind' of system (atom) are introduced without operational definition; they happen to correspond to actual situations in microphysical experiments, however.

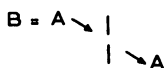


Fig. 1a

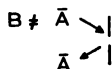


Fig. 1b

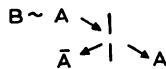


Fig. 1c

As an illustration, A may signify a state of vertical orientation of the molecular axis of a certain kind of particle, and the A -filter may be a screen with a vertical slit. State \bar{A} may be a state of horizontal orientation of the same particle, so that the A -filter blocks \bar{A} -state particles.

Imagine now that, starting from a state $B \neq A$ (Fig. 1b) one gradually 'changes' state B so that it becomes 'more similar' to A (again no operational definition of the terms in quotation marks is given). One may expect *a priori* that an abrupt change from Fig. 1b to 1a will take place only in the last moment when B becomes exactly equal to A . The *postulate of continuity of cause and effect* requires, however, that a gradual change from $B \neq A$ to $B = A$ as cause will lead to a gradual change of effect, from all B 's blocked to all B 's passed by the A -filter. More precisely, the continuity postulate requires that there be intermediate states B between $B \neq A$ and $B = A$, with results intermediate between Fig. 1b and 1a, that is, with some B 's passing and some rejected, as pictured in Fig. 1c; such cases then signify a 'fractional equality' between B and A , written $B \sim A$. The ratio between passed and repelled B -state particles can only be a *statistical ratio*, i.e. a probability ratio for an individual B -state particle. *Individual indeterminacy controlled by statistical ratios is a consequence of the continuity postulate* for cause and effect. The passing fraction written $P(B, A)$ of B -state particles through the A -passing filter may be taken as an operational definition of the *fractional equality degree* between the states A and B , of value between 0 and 1. And since equality degrees ought to be mutual, one will introduce the *symmetry postulate*, $P(A, B) = P(B, A)$; the latter is physically justified as the statistical counterpart of the reversibility of deterministic processes. It stipulates that the statistical fraction of B -state particles passed by an A -passing

filter equals the statistical fraction of A -state particles passing a B -filter.

Similar considerations apply to any game of chance with the alternative 'yes' or 'no', passed or blocked, right of left, etc. For example, when balls are dropped from a chute upon a knife edge, they will drop to he right or to the left, depending on the aim of the chute.

According to the continuity postulate, however, there ought to be a continuity of cases between all balls to the right and all to the left, occurring within a small range of physical aim, with statistically ruled ratios of r - and l -balls, gradually changing from 100:0 to 0:100 when the physically regulated aim of the chute is changed from one to the other end of the small angular range. Hypothetical reservations about concealed causes for individual r - and l -events would never explain the miracle of statistical cooperation' of individual events yielding fixed statistical ratios [1], [2].

Next we introduce the empirical *postulate of reproducibility of a test result* which stipulates that a B -state particle in Fig. 1c which has once passed the A -filter will pass an A -filter again with certainty. This harmless looking postulate implies that the incident B -state particle, in the first act of passing the A -filter, must have changed its state from B to A . Indeed, only thus will it pass another A -filter again with certainty. Similarly, an incident B -state particle once repelled by the A -filter must have jumped, by virtue of its first repulsion, from B to the new state \bar{A} so that it will be repelled again if tested once more by the A -filter. Discontinuous changes of state (transitions, jumps) in reaction to a testing instrument can thus be seen as consequences of the postulate of *reproducibility* of a test result and *continuity* of cause and effect. To these postulates we have added that of symmetry, $P(A, B) = P(B, A)$, in which P now assumes the meaning of a *transition probability* from state B to A in an A -filtertest, and from A to B in a B -filtertest.

2. The Probability Schema. After these ideological preparations we come to the mathematical schema of the probabilities of transition. Consider a class of entities S (= 'states' of a given atom) which are in a mutual relation of 'fractional equality' $S_m \sim S_n$, quantitatively described by positive fractional numbers, $P(S_m, S_n)$, denoted as 'equality fractions'. Special cases are $P = 0$ (separability, total inequality of S_m and S_n) and $P = 1$ (identity, inseparability). The P -relations permit a division of the elements of class S into subclasses, the subclass A with members $A_1 A_2 \dots$ which satisfy the *orthogonality relation*

$$(1) \quad P(A_m A_{m'}) = \delta_{mm'},$$

the subclass B , and C , and so forth. (The selection of complete orthogonal subclasses out of the entirety of entities S is not unique, a fact known to the quantum theorist as 'degeneracy').

P -values connecting the elements of two subclasses such as A and B may be arranged in a matrix:

$$(2) \quad (P_{AB}) = \begin{Bmatrix} P(A_1, B_1) & P(A_1, B_2) & \dots \\ P(A_2, B_1) & P(A_2, B_2) & \dots \\ \dots & \dots & \dots \end{Bmatrix}$$

The physical interpretation of the P 's as probabilities of transition in tests justifies the postulate that the sum of the transition probabilities from any one state A_m to the various states $B_1 B_2 \dots$ be unity, i.e. that each row of the matrix (2) sums up to unity. Furthermore, according to the *symmetry postulate*

$$(3) \quad P(A_m, B_n) = P(B_n, A_m),$$

the columns of the matrix (P_{AB}) are the rows of the matrix (P_{BA}) so that the columns of (P_{AB}) also have sum unity;

$$(3') \quad \sum_n P(A_m, B_n) = 1 \text{ and } \sum_m P(A_m, B_n) = 1.$$

Suppose now that the matrix (2) has M rows and N columns. The sum of all its elements would then be M when summing the rows, and N when summing the columns. Thus $M = N$, that is, the matrices (P_{AB}) and (P_{AC}) etc. must be *quadratic*, and the subclasses A, B, C, \dots must all have the *same multiplicity*, M . The multiplicity M of the orthogonal sets of states may be finite or infinite depending on the 'kind' of particle. The P -matrices are *unit magic squares*.

3. The Probability Metric. We now introduce the further postulate that the various P -matrices are interdependent by virtue of a *general law* according to which one matrix (P) in a group is determined by the other matrices (P) of the same group. Only the following simple interdependence laws between two-index quantities are feasible:

$$(4) \text{ the addition law } U_{AC} = U_{AB} + U_{BC} \\ \text{made self-consistent by } U_{AB} = -U_{BA}$$

and corresponding laws for distorted quantities $W = f(U)$, e.g. for $W = e^U$.

(5) the *multiplication law* $W_{AC} = W_{AB} \cdot W_{BC}$

made self-consistent by $W_{AB} \cdot W_{BA} = 1$

There is no other conceivable way of making U_{AC} or W_{AC} independent of the choice of the intermediate entity B than the addition theorem (4) and its generalization by distortion.

A model of (4) is furnished by the geometry of lengths L_{AB} , L_{AC} , etc. in frameworks connecting points A , B , C , Although (4) cannot be applied to the lengths L themselves, it may be applied to a substructure of quantities φ satisfying the triangular relation $\varphi_{AC} = \varphi_{AB} + \varphi_{BC}$ with $\varphi_{AB} = -\varphi_{BA}$, known as *vectors*. The latter determine the lengths $L = |\varphi|$. Of particular interest is *plane* geometry where vectors φ can be written as complex symbols, $\varphi = |\varphi| \cdot e^{i\alpha}$. Also in a plane, 5 points are connected by 10 lengths; when 9 of them are given they uniquely determine the tenth L .

In order to construct a law of interdependence between unit magic squares one may start from (5). Although (5) cannot be applied to the matrices (P) themselves, it may be applied to a substructure of quantities ψ which are to satisfy the matrix multiplication formula

$$(6) \quad (\psi_{AC}) = (\psi_{AB}) \cdot (\psi_{BC}), \text{ with } (\psi_{AA}) = (\psi_{AB}) \cdot (\psi_{BA}) = (1).$$

When now decreeing (the asterisk standing for the complex conjugate):

$$(7) \quad \psi(A_k, B_n) = \psi^*(B_n, A_k) \text{ and } P = |\psi|^2,$$

the P -matrices become unit magic squares, as required. (6) is known as the law of *unitary transformation*, connecting 'orthogonal axes systems' A and B etc. by 'complex directional cosines' ψ . A tensor f in general obeys the transformation formula

$$(8) \quad (f_{AD}) = (\psi_{AB}) \cdot (f_{BC}) \cdot (\psi_{CD}).$$

To the physicist, the quantities ψ are the 'probability amplitudes' which satisfy the *law of interference* (6), and the tensors f are 'observables'. When f has its eigenvalues in the states $F_1 F_2 \dots$ that is, when

$$(9) \quad f(F_n, F_{n'}) = f(F_n) \cdot \delta_{nn'},$$

then, as a special case of (8), one has

$$(9') \quad f(A_k, A_j) = \sum_n \psi(A_k, F_n) \cdot f(F_n) \cdot \psi(F_n, A_j).$$

The ψ -interference law and the corresponding transformation law for

observables was first found inductively and was considered as a most surprising empirical law of nature. It turns out to be the only conceivable solution of the mathematical problem of finding a general self-consistent law connecting *unit magic squares*, viz. the law of unitary transformation.

In opposition to numerous physicists who see in the interference law for complex probability amplitudes a profound and unfathomable plan of nature presenting us with an abstract and unpictorial substructure of reality manifest in a wave-particle duality, it may be noticed that (a) each complex ψ may be pictured as a vector in a plane giving direction to the corresponding probability P so that the P -metric can be visualized as a structural framework of lines in a plane, and (b) similar to plane geometry where 5 points A, B, C, D, E are connected by 10 lengths L_{AB}, L_{AC} , etc. and 9 L 's uniquely determine the tenth L , so are there direct relations between the 10 unit magic square matrices $(P_{AB}), (P_{AC})$, etc. which connect 5 orthogonal sets of states so that 9 P -matrices uniquely determine the tenth. That is, there are direct relations between the real probabilities P which can be formulated without resorting to complex quantities ψ with wave-like phase angles.

4. Quantum Periodicity Rules. The quantum theorems of Born and Schrödinger

$$(10) \quad (qp - pq) = \hbar/2i\pi \text{ and } p = (\hbar/2i\pi)\partial/\partial q$$

are equivalent to the rule that the amplitude function $\psi(q, p)$ is a complex exponential function

$$(11) \quad \psi(q, p) = \exp(iqp/\text{const})$$

with $\text{const} = \hbar/2\pi$. The quantum rules (10) or (11) are usually introduced *ad hoc* as inductive results of quantum experience. I am going to show that they are consequences of the following postulates added to those introduced before:

- a) Linear coordinates q and linear momenta p are physically defined up to additional constants so that there are observables whose values depend on q -differences and on p -differences only.
- b) The statistical density of conjugates q and p is *constant* in qp -space (as it is in classical statistical mechanics).

The proof of (11) on the grounds of (a)(b) rests on the fact that the complex exponential function, $f(x) = \exp(ix/\text{const})$ is the only function $f(x)$ which, together with its complex conjugate $f^*(x)$, satisfies the condi-

tion that the product $f(x_1) \cdot f^*(x_2)$ will depend on the *difference* $x_1 - x_2$ only.

The detailed proof runs as follows. As a special case of (9') for an observable f defined as a function of q one has

$$f(p_k, p_j) = \sum_n \psi(p_k, q_n) f(q_n) \psi^*(q_n, p_j).$$

If q is a linear coordinate running continuously from $-\infty$ to $+\infty$, and for given p -values has *constant* $|\psi|^2$ density, the last formula becomes an integral with *constant* weight factor in the integrand:

$$(12) \quad f(p_k, p_j) = \int \psi(p_k, q) f(q) \psi^*(q, p_j) dq$$

Since $f(q)$ may be any observable whatsoever, one may consider the case that it is a δ -function with maximum at any chosen place q_i ; the integral then reduces to

$$f(p_k, p_j) = \psi(p_k, q_i) \psi^*(p_j, q_i).$$

If the 'transition value' $f(p_k, p_j)$ is to depend on the *difference* $p_k - p_j$ only, the function ψ on the right must contain p in the form

$$(13) \quad \psi(q, p) = \exp(\dots i p \dots).$$

An analogous consideration applied to an observable $g(p)$ which may be chosen as a δ -function yields the result that the function Ψ must contain q in the form

$$(13') \quad \psi(q, p) = \exp(\dots i q \dots).$$

(13) and (13') together leave only the following alternative: *Either* $\psi(q, p)$ is of the form

$$\psi(q, p) = \exp(\alpha i q + \beta i p)$$

with separate real constant factors α and β , or

$$(14) \quad \psi(q, p) = \exp(i \gamma q p)$$

with common real factor γ . The first alternative would lead, according to (12) to

$$f(p_k, q_j) = \exp(i \alpha (p_k - p_j)) \int f(q) \cdot dq = \exp(i \alpha (p_k - p_j)) \cdot \text{const},$$

where the left hand side depends on the choice of the function f , whereas the right hand side does not. Only the second alternative makes sense. When writing $\hbar/2\pi$ for γ Eq. (14) it is identical with (11), q.e.d. Eq. (11) is

the fundamental wave function of quantum dynamics with wave length $\lambda = h/p$.

For completeness sake we add the well-known deduction of the symmetry theorems which are of such decisive importance for the aggregation of identical particles. Identity of two particles a and b signifies their indiscernibility and in particular equality of the two transition probabilities

$$P(S; a_i, b_j) = P(S; b_i, a_j),$$

or omitting reference to S :

$$|\psi(a_i, b_j)|^2 = |\psi(b_i, a_j)|^2.$$

This equation can be satisfied only when ψ is either symmetric or anti-symmetric with respect to an exchange of the letters a and b , proved as follows. Write

$$\begin{aligned}\psi(a_i, b_j) &= \frac{1}{2}[\psi(a_i, b_j) + \psi(b_i, a_j)] + \frac{1}{2}[\psi(a_i, b_j) - \psi(b_i, a_j)] \\ &= \phi_{\text{sym}}(a, b) + \phi_{\text{ant}}(a, b).\end{aligned}$$

Similarly

$$\psi(b_i, a_j) = \phi_{\text{sym}}(a, b) - \phi_{\text{ant}}(a, b)$$

Taking the absolute squares of the two last equations one arrives at

$$P(a_i, b_j) = |\phi_{\text{sym}}|^2 + |\phi_{\text{ant}}|^2 + \text{real part of } (2\phi_{\text{sym}}\phi_{\text{ant}}^*)$$

$$P(b_i, a_j) = \text{same real part of same}$$

The two P 's can be equal only when either ϕ_{sym} or ϕ_{ant} vanishes, i.e. (excluding the trivial case of $\psi \equiv 0$) when either $\psi = \phi_{\text{ant}}$ or $\psi = \phi_{\text{sym}}$, q.e.d.

For systems of three or more identical particles $\psi(a, b, c, \dots)$ must either be symmetric with respect of the exchange of each pair, or anti-symmetric. Indeed, if ψ were symmetric with respect to a and b , but antisymmetric with respect to a and c , one would arrive at the following sequence:

$$\begin{aligned}+ \psi(a, b, c) &= + \psi(b, a, c) = - \psi(b, c, a) = - \psi(a, c, b) = \\ &= + \psi(c, a, b) = + \psi(c, b, a) = - \psi(a, b, c)\end{aligned}$$

which is self-contradictory. All particles are thus divided in two classes,

those which form symmetric, and those which form antisymmetric ψ -functions.

This concludes the deduction of the quantum theorems from basic postulates of a non-quantal character.

5. Quantum Fact and Fiction. A few remarks may be added concerning the present quantum philosophy, reputedly the most revolutionary innovation in the theory of knowledge of the century. Its starting point is the allegation that quantum theory has invalidated the notion of *objective states* possessed by a microphysical system independent of an observer (according to some authorities) or independent of a measuring instrument (according to others). And the quantity ψ is said to have a particularly 'subjective' character in so far as it expresses expectations of an observer, rather than states of an atom. ψ is also reputed to be 'abstract' and 'unanschaulich' (unpictorial) due to its complex-imaginary form.

In the writers opinion, this quantum philosophy rests on various misunderstandings and fictions. *First*, complex quantities stand for vectors in a plane; hence ψ gives direction to the transition probabilities so that the latter form a structural framework in a plane. The ψ -multiplication law (6) is quite analogous to the geometrical vector addition law $\varphi_{AC} = \varphi_{AB} + \varphi_{BC}$. But nobody has yet found plane geometry abstract and unpictorial because it connects real lengths by vectors which could be symbolized by complex numbers.

Second, since a test resulting in the state A_m of an atom is *reproducible* by means of the same A -meter, one may legitimately denote the state A_m as being 'objectively possessed' by the atom. It is true that a subsequent B -test throws the atom into a new (equally reproducible) state B_n . Thus one does not have the right to say, or even to imagine, that the atom is in the two states A_m and B_n simultaneously; the two states are 'incompatible'. But incompatibility as such is nothing novel and revolutionary. A state of angular twist value w of a rod of ice, and a viscosity value v of the same sample in the liquid state are mutually incompatible; there are no combination wv -states. It is significant of quantum dynamics that a state q and a state p , though individually reproducible, do not allow reproducible 'objective' qp -states; and if an objective q -state has been ascertained one must not even *imagine* any hidden simultaneous p -value to prevail. But this is not initiating a new philosophy of knowledge. It merely tells us to be careful with the application of the term 'objective state'. Of course, physicists are more impressed by the example of qp -

incompatibility than by the trivial example of *vw*-incompatibility. Yet after thirty years of emphasizing differences, one may as well begin stressing similarities between quantum physics and everyday experience.

Third, in this connection one ought to remember that statistical law, as opposed to classical determinism, is known from ordinary games of chance; they, too, confront us with the 'miracle of statistical cooperation' of individual events irreducible *in principle* [1], [2] to hidden causes. There is no structural difference between the ordinary ball-knife game described above and the quantum game of Fig. 1c.

Fourth a great issue has been made of ψ being a subjective expectation function which suddenly collapses or contracts in violation of the 'wave equation' when a definite observation is made, turning potentiality into actuality. However, in spite of subjectively tainted words 'expectation' and 'probability', the quantum theory, like any other theory in physics, correlates experimental data rather than mental states; in particular it correlates statistical experience gained in tests of atoms with macroscopic instruments. If someone uses these statistical laws (which are of the same quality as the Gauss law of errors) for placing bets or for enjoying anticipations of future events, this is his personal affair and has nothing to do with the quantum theory. (Similarly, nobody has yet found a subjective element in Gauss' error law, or in Newton's law of attraction because astronomers anticipate eclipses with high accuracy). The fiction that quantum theory deals with differential equations for expectations rather than with the correlation of objective data which never collapse, has instilled utter confusion into the 'quantum theory of measurement'. Here we learn that a ψ -function, after first developing according to the Schrödinger equation as a kind of 'process equation of motion', suddenly collapses whenever a point event takes place (according to some authorities) or only when an observer takes notice of the point event (according to others). But since nobody can seriously believe in such inconsistencies, one tries at least to talk away the difficulty, as testified by extended discussions at many symposiums on 'measurement' during the last thirty years. The chief trouble is the mistaken view that the Schrödinger equation describes a physical change of state, either individually or statistically. Actually if connects various mathematical 'representations' of one and the same fixed state with one another, be it the fixed state *A* before the measurement, or *B* after the measurement [3], [4], [5], [6].

Fifth, confusion prevails also with respect to the famous waveparticle *duality*. In fact the latter has become illusory since Max Born thirty

years ago introduced the statistical particle interpretation of the 'wave function' and thereby restored a *unitary particle theory*, following a short period of doubt whether matter really consisted of waves or of particles. Before Born it was considered philosophical to argue that neither waves nor particles are 'real'; but the same pseudo-philosophical *talk* has survived although physicists in their sober hours consider particles, and particles alone, as the constituting substance of matter (in the non-relativistic domain). Still talking of duality, i.e. drawing a parallel between a thing (particle) and one of its many qualities (its occasional periodic probability distribution in space and time) is illogical.

The great merit of Schrödinger's original matter wave theory had been that it gave an explanation of the discreteness of quantum states in terms of proper vibrations in a medium. But Born's statistical interpretation, confirmed by the observation of point events, destroyed the explanatory character of the Schrödinger waves, without substituting a rational explanation for the wave-like phenomena. The present investigation is to fill this gap. The wave-like ψ -interference becomes a natural and necessary quality of particles under the postulate that the unit magic square P-tables are connected by a self-consistent law, the only conceivable such law is that of unitary transformation, which is identical with that of ψ -interference (6). Furthermore, the wave-like qp -periodicity, the basis of all 'quantization', becomes a natural and obvious particle quality under the postulates (a)(b) for conjugate observables q and p .

Postscript: The deduction on p. 360 is inconclusive. Only perturbation theory leads to the symmetry principles.

Bibliography

- [1] LANDÉ, A., *The case for indeterminism*. In 'Determinism and Freedom', edited by Sidney Hook, New York University Press (1958), p. 69.
- [2] —, *Determinism versus continuity in modern science*. *Mind*, vol. 67 (1958), pp. 174–181.
- [3] —, *Foundations of quantum theory*. Yale University Press, 1955.
- [4] —, *The logic of quanta*. *British Journal for the Philosophy of Science*, vol. 6 (1956), pp. 300–320.
- [5] —, *Non-quantal foundations of quantum theory*. *Philosophy of Science*, vol. 24 (1957), pp. 309–320.
- [6] —, *Zeitschrift für Physik*, vol. 153 (1959) pp. 389–393.

QUANTENLOGIK UND DAS KOMMUTATIVE GESETZ

PASCUAL JORDAN

Universität Hamburg, Hamburg, Deutschland

In bekannter Weise arbeitet die Quantenmechanik mit Operatoren oder Matrizen; und wenn wir uns die Grundgedanken der Quantenmechanik klar machen wollen, so ist es empfehlenswert, daß wir die mathematischen Probleme, die mit der Theorie *unendlicher* Matrizen zusammenhängen, ganz ausschalten. Wir haben es dann, mathematisch gesprochen, nur mit *Algebra* zu tun.

Wir denken uns also ein quantenphysikalisches System (Beispiele wären leicht zu nennen), dessen meßbare Eigenschaften darzustellen sind durch die Matrizen eines endlichen Grades n , wobei die Matrixelemente beliebige komplexe Zahlen sind. Die Theorie lehrt bekanntlich:

Jeder *hermiteschen* Matrix A innerhalb der Algebra dieser Matrizen entspricht eine meßbare Größe (anders gesagt: eine mögliche Struktur eines auf das System anwendbaren Meßinstrumentes). Die *Eigenwerte* der Matrix A sind die möglichen Meßresultate, die sich bei Messung von A ergeben können. Mathematisch ist ja die Matrix A darstellbar in der Form

$$A = \sum_k \alpha_k e_k, \quad (1)$$

wobei die e_k orthogonale (hermitesche) Idempotente sind: $e_k e_j = \delta_{kj} e_k$, während die α_k die Eigenwerte von A bedeuten. Die Aussage: Als Meßergebnis an der Größe A hat sich der Eigenwert α_1 ergeben, kann also ersetzt werden durch die Aussage, daß eine Messung der Größe e_1 für diese ihren Eigenwert 1 (und nicht ihren anderen Eigenwert 0) ergeben hat. Wir brauchen also nur von den Idempotenten zu sprechen.

Die idempotente Größe e_1 , bei deren Messung der Eigenwert 1 gefunden wurde, sei insbesondere *unzerlegbar*, also nicht als Summe von zwei orthogonalen Idempotenten darstellbar. Dann werde nachfolgend ein beliebiges Idempotent e' gemessen. Wie groß ist die Wahrscheinlichkeit, daß wir für e' den Wert 1 finden? Die Quantenmechanik (oder die „statistische Transformationstheorie“) antwortet:

$$w = Sp(e_1 e'). \quad (2)$$

Mit Sp ist die *Spur* der Matrix $e_1 e'$ gemeint.

In diesen Formulierungen ist der ganze grundsätzliche Inhalt der Quantenmechanik zusammen gefaßt.

Man kann aber der Sache eine andere Fassung geben, welche mit der soeben erläuterten mathematisch äquivalent ist. Wir betrachten eine projektive Geometrie von $n - 1$ Dimensionen, oder anders ausgedrückt, wir betrachten Einheitsvektoren in einem Raum von n Dimensionen. Die Komponenten ξ_k solcher Vektoren sollen beliebige komplexe Zahlen sein. Jedes unzerlegbare Idempotent ist dann darstellbar als Matrix

$$e' = (\xi_k^* \xi_1) \text{ mit } Sp(e') = \sum_k |\xi_k|^2 = 1. \quad (3)$$

(Mit ξ^* bezeichnen wir die Konjugierte zu ξ). Allgemeiner besteht umkehrbar eindeutige Zuordnung zwischen den linearen Scharen der betrachteten Vektoren (oder den linearen Unterräumen der projektiven Geometrie) und den hermiteschen Idempotenten der früher betrachteten Matrixalgebra. Wir können also, statt von den Idempotenten zu sprechen, von den zugehörigen Vektorscharen sprechen. Sind in (2) beide Idempotenten unzerlegbar, so haben wir (in unmittelbar verständlicher Bezeichnungsweise)

$$w = Sp(e_1 e') = |\sum_k \xi_k^* \eta_k|^2. \quad (4)$$

Diese zweite Formulierungsweise der quantenmechanischen Grundgesetze lehnt sich enger als die andere an die *Schrödingersche* „Wellenmechanik“ an.

Es ist aber eine *dritte*, nochmals anders ausschende Formulierungsweise möglich, die von *Birkhoff* und *v. Neumann* vorgetragen worden ist. Die $(n - 1)$ -dimensionale projektive Geometrie kann mathematisch erklärt werden als ein *Verband* („lattice“) von bestimmten Eigenschaften; und die dadurch ermöglichte Einordnung der Quantenmechanik in die mathematische Theorie der Verbände gibt uns einen überraschenden neuen Einblick: Wir können danach den Übergang von der klassischen Mechanik zur Quantenmechanik — der ja gewöhnlich als Übergang von *kommutativer* zu *nichtkommutativer Algebra* der meßbaren Größen betrachtet wird — auch als einen Übergang von *distributiven Verbänden* zu *modularen Verbänden* auffassen.

In der klassischen Mechanik können wir jede durch ein Meßergebnis begründete *Information* oder *Aussage* über den Zustand eines Systems so ausdrücken, daß der den Zustand des Systems darstellende Punkt im *Phasenraum* sich innerhalb einer gewissen Punktmenge *a* des Phasen-

raums befindet. Der Durchschnitt $a \cap b$ zweier Punktmengen im Phasenraum entspricht also der Verknüpfung der beiden zugehörigen Aussagen durch „und“; die Vereinigungsmenge $a \cup b$ entspricht der Verknüpfung beider Aussagen durch „oder“. In dieser Weise ist unmittelbar ersichtlich, daß die Gesamtheit der möglichen *Aussagen* über den Zustand des klassischen mechanischen Systems ebenso wie die Teilmengen einer Punktmenge einen *distributiven Verband* bilden. Dabei entspricht ferner der *Verneinung* einer Aussage der Übergang von der Punktmenge a zur komplementären Punktmenge \bar{a} .

Bekanntlich sprechen wir in der Mathematik von einem Verband, wenn für eine gewisse Elementenmenge a, b, \dots Verknüpfungen \cap, \cup definiert sind, welche *assoziativ* und *kommutativ* sind, und außerdem das Axiom

$$(a \cap b) \cup a = a \cap (b \cup a) = a \quad (5)$$

erfüllen, aus welchem die Idempotenz aller Elemente für diese beiden Verknüpfungen folgt:

$$a \cap a = a \cup a = a. \quad (6)$$

Besteht zwischen zwei speziellen Elementen a, b die Beziehung $a \cap b = a$ (äquivalent mit $a \cup b = b$), so schreiben wir auch $a \subseteq b$; diese Beziehung des *Enthaltenseins* ist *reflexiv* und *transitiv*. Aus $a \subseteq b$ und $b \subseteq a$ folgt $a = b$.

Man nennt bekanntlich einen Verband *distributiv*, wenn er das zusätzliche Axiom

$$a \cap (b \cup c) = (a \cap b) \cup (a \cap c) \quad (7)$$

erfüllt. Es gilt dann gleichzeitig auch das dazu *duale*, aus (7) durch Vertauschung der Zeichen \cap, \cup entstehende Gesetz; eine Tatsache, die man z.B. so beweisen kann, daß man (7) als äquivalent mit folgender dualsymmetrischer Beziehung erweist:

$$(a \cap b) \cup (a \cap c) \cup (b \cap c) = (a \cup b) \cap (a \cup c) \cap (b \cup c). \quad (8)$$

Endlich sei erwähnt, daß für den Übergang von einer Teilmenge a zur komplementären \bar{a} („*Verneinung*“) folgende Axiome gelten:

$$\left. \begin{aligned} \bar{\bar{a}} &= a; \quad \overline{a \cap b} = \bar{b} \cup \bar{a} \\ a \cap \bar{a} &= 0 = \text{leere Menge}; \quad a \cup \bar{a} = 1 = \text{volle Menge} \end{aligned} \right\} \quad (9)$$

Betrachten wir nun statt eines klassischen Systems ein quantenmechanisches (wiederum mit endlichem Grad n seiner Matrixalgebra), so treten

an die Stelle von Punktmengen in Phasenraum die hermiteschen Idempotenten oder die ihnen umkehrbar eindeutig zugeordneten linearen Unterräume der $(n - 1)$ -dimensionalen projektiven Geometrie. Diese erlauben ebenfalls Verknüpfungen \cap , \cup , nämlich im Sinne des Durchschnitts $a \cap b$ von a und b , sowie des durch a und b aufgespannten linearen Raumes $a \cup b$. Aber der damit definierte Verband ist nicht mehr distributiv, sondern erfüllt statt dessen nur noch das schwächere *Dedekindsche Modularaxiom*, welches — um sogleich seine ebenfalls dualsymmetrische Bedeutung zu zeigen — folgendermaßen formuliert werden kann:

$$(a \cap b) \cup [c \cap (a \cup b)] = [(a \cap b) \cup c] \cap (a \cup b). \quad (10)$$

Man kann diesen Umstand nach *Birkhoff-Neumann* so ausdrücken, daß man von einer *Quantenlogik* im Gegensatz zu einer *klassischen Logik* spricht. Natürlich ist es Geschmacksache, ob man diese Bezeichnung anerkennen will; jedoch ist sie jedenfalls dann naturgemäß, wenn man unter „Logik“ die Gesetze der möglichen Verknüpfungen von Aussagen oder Informationen über den Zustand eines physikalischen Systems verstehen will — in *dieser* Auffassungsweise ist auch die Logik eine empirische Wissenschaft, weil nur empirisch klargestellt werden kann, welche Gesamtheit möglicher Aussagen zu einem bestimmten physikalischen System hinzugehört. (Offenbar ist es keineswegs im Widerspruch hierzu, daß man andererseits alle auf die Quantentheorie bezüglichen Überlegungen unter alleiniger Verwendung der klassischen, also distributiven Logik formulieren und durchführen kann.)

Es gibt auch in der Quantenlogik eine *Verneinung*, nämlich $\bar{e} = 1 - e$, für welche die Axiome (9) gelten, wobei jetzt 0 und 1 als die durch diese Zeichen bezeichneten Elemente der Matrixalgebra zu verstehen sind.

Entscheidend für die Rechtfertigung der *Birkhoff-Neumannschen* Betrachtungsweise ist aber folgender von *Neumann* aufgestellter mathematischer SATZ: *Ein modularer Verband, welcher einige zusätzliche Eigenschaften hat (er muß irreduzibel sein, nur endliche Ketten des Enthaltenseins zulassen, und „komplementierbar“ sein), ist immer eine projektive Geometrie* endlicher Dimension. „Komplementierbar“ ist er insbesondere dann — in einer spezielleren Weise — wenn es in ihm auch eine Operation der Verneinung in besprochener Form gibt. In diesem Falle hat der zu der projektiven Geometrie gehörige Schiefkörper insbesondere die Eigenschaft, welche ich mit dem Wort „*formal komplex*“ bezeichnet habe. Soll dieser Schiefkörper den reellen Zahlkörper in sich enthalten, so muß er

entweder dieser selbst sein oder der Körper der komplexen Zahlen oder der Schiefkörper der Quaternionen.

Ein reizvoller *Neumannscher* SATZ besagt übrigens, daß die modularen Verbände durch folgende Eigenschaft gekennzeichnet sind: *Gilt für drei spezielle Elemente a, b, c die Distributiv-Beziehung (7), so ist sie invariant gegen Permutationen dieser drei Elemente.* Weitergehend kann man zeigen (*Jordan*), daß dann der ganze durch a, b, c erzeugte Teilverband distributiv ist; und das erlaubt folgende Klarstellung: *Zwei Idempotente e, e' sind genau dann vertauschbar, also $ee' = e'e$, wenn zwischen e, \bar{e}, e' eine Distributivbeziehung besteht.*

Nach diesen Vorbereitungen komme ich zur Besprechung eines Gedankens, der mich zeit langer Zeit beschäftigt hat. Für die Weiterentwicklung der Quantentheorie könnte es notwendig werden, den grundsätzlichen Formalismus der Quantenmechanik, wie er besprochen wurde, zu erweitern oder zu verallgemeinern. Gibt es dazu mathematische Möglichkeiten?

Diese Frage ist zunächst in der Weise untersucht worden, daß Verallgemeinerungen der assoziativen Matrix-Algebren untersucht worden sind [1, 2, 4]. Diese Untersuchungen haben Anlaß zu einer ganzen Reihe weiterer mathematischer Untersuchungen gegeben [5]. Jedoch soll diese Seite der Entwicklung jetzt nicht ausführlicher besprochen werden, da sie trotz mancher reizvoller mathematischer Ergebnisse für die Physik bislang nichts Fruchtbare ergeben hat.

Es liegt aber nahe, eine andere Verallgemeinerungsmöglichkeit zu studieren, darin bestehend, daß man innerhalb der Quantenlogik noch einmal den Übergang *vom Kommutativen zum Nichtkommutativen* versucht. In der Tat hat sich gezeigt, daß die Theorie der Verbände sich durch Verzicht auf das Axiom der Kommutativität zu einer Theorie der „*Schrägverbände*“ (*skew lattices*) verallgemeinern läßt, welche zwar eine Fülle neuer, zum Teil recht schwieriger Fragen aufwirft, aber auch viele schöne Ergebnisse schon jetzt ermöglicht hat, von denen im Folgenden nur eine kurze Andeutung gegeben werden kann. Diese nichtkommutative Verallgemeinerung der Verbandstheorie ist zuerst von *Klein-Barmen* ins Auge gefaßt, später vom Verfasser in Angriff genommen, und unabhängig davon auch von *Matsushita* (vergleiche [3]). Die Überlegungen des Verfassers sind durch die Mitarbeit von *E. Witt* und *W. Böge* entscheidend gefördert worden.

Wir denken uns eine Elementenmenge mit zwei *assoziativen* Verknüp-

fungen \vee , \wedge . Wir fordern ferner als Grundaxiom

$$(a \wedge b) \vee a = a \wedge (\vee a) = a, \quad (11)$$

woraus auch jetzt die Idempotenz

$$a \wedge a = a \vee a = a \quad (12)$$

folgt. Während aber in (7) das kommutative Gesetz mannigfache Umstellungen der Buchstaben zuläßt, sollen die dadurch entstehenden Formeln keineswegs auch auf die Schrägverbände übertragen werden. Beispielsweise wird das — von (11) unabhängige — zusätzliche Axiom

$$(b \wedge a) \vee a = a \wedge (a \vee b) = a \quad (13)$$

nur von einer sehr speziellen, ziemlich trivialen Klasse von Schrägverbänden erfüllt.

Es gibt nun in jedem Schrägverband *vier* Formen eines reflexiven und transitiven Enthaltenseins, die im allgemeinen verschiedene Bedeutung haben — sind sie in einem speziellen Schrägverband alle vier gleichbedeutend, so ist dieser kommutativ, also ein Verband. Im allgemeinen Falle kommen auch entsprechende Äquivalenzklassen von mehr als einem Element vor. Die vier Formen des Enthaltenseins von a in b sind definiert durch die Beziehungen:

	links	rechts	
stark	$b \wedge a = a$	$b \vee a = b$	(14)
schwach	$a \vee b = b$	$a \wedge b = a$	

Jede Form des starken Enthaltenseins ergibt als Folgerung die zugehörige Form schwachen Enthaltenseins, was wir so andeuten können:

↓	↓

(11')

Das *zusätzliche Axiom* (13) ist dann *gleichbedeutend* damit, daß beide Formen schwachen Enthaltenseins *stets zugleich* vorliegen:

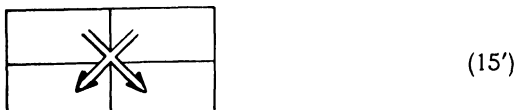
↔	↔

(13')

Schwächer als (13) ist das Axiom

$$\left. \begin{aligned} a \wedge b \wedge a &= a \wedge b, \\ a \vee b \vee a &= b \vee a, \end{aligned} \right\} \quad (15)$$

welches *gleichbedeutend* ist mit folgenden Beziehungen hinsichtlich des Enthaltenseins:



Offenbar bekommen wir (15') als eine Folgerung aus (11') und (13').

Da für die Quantentheorie nicht alle beliebigen Verbände, sondern nur *modulare* Verbände von Bedeutung sind, so scheint folgende Tatsache ermutigend — welche unabhängig von physikalischen Spekulationen auch rein mathematisch reizvoll ist: *Man kann den Begriff „modular“ auf die Schrägverbände in einfacher und schöner Weise übertragen.* Nämlich in Gestalt des folgenden Axioms, welches genau der Formel (10) nachgebildet ist — es kommt jetzt aber entscheidend auf die Reihenfolge der Zeichen an, welche in (10) aufgrund der Kommutativität weitgehend beliebig war:

$$(a \wedge b) \vee [c \wedge (a \vee b)] = [(a \wedge b) \vee c] \wedge (a \vee b). \quad (16)$$

Die Analogie zum kommutativen Fall bewährt sich dabei auch in folgendem Sinne: Man kann die Formel (10) ersetzen durch das damit gleichwertige Axiom, daß $x \subseteq y$ stets die Folgerung

$$x \cup (c \cap y) = (x \cup c) \cap y \quad (17)$$

haben soll. Ganz entsprechend ist (16) äquivalent mit folgender Aussage: *Ist x zweifach schwach enthalten in y , so gilt*

$$x \vee (c \wedge y) = (x \vee c) \wedge y. \quad (18)$$

Auch erweist sich der durch (16) definierte Begriff der „modularen Schrägverbände“ darin als sinnvoll und angemessen, daß es tatsächlich eine große Fülle von Beispielen für diesen Begriff gibt.

Zur Konstruktion weiter Klassen von Beispielen von Schrägverbänden ist folgendes Verfahren geeignet: Angenommen, es sei uns ein gewisser Schrägverband \mathfrak{B} bereits gegeben; es kann sich insbesondere um einen kommutativen, also einen Verband handeln. Wir wollen die Verknüp-

fungen innerhalb von \mathfrak{B} mit den Zeichen \cap , \cup bezeichnen; dann aber definieren wir in \mathfrak{B} neue Verknüpfungen \wedge , \vee durch

$$\left. \begin{aligned} a \vee b &= fa \cup fb, \\ a \wedge b &= a \cap Fb. \end{aligned} \right\} \quad (19)$$

Hierbei sollen die Elemente fx bzw. Fx von \mathfrak{B} gewisse *Funktionen* des Elementes $x \in \mathfrak{B}$ bedeuten; und zwar mögen diese Funktionen folgende Eigenschaften haben:

$$\left. \begin{aligned} f(fa \cup b) &= fa \cup fb, \\ fa \cup a &= a; \\ F(a \cap Fb) &= Fa \cap Fb, \\ a \cap Fa &= a. \end{aligned} \right\} \quad (20)$$

Dann ist die Elementenmenge \mathfrak{B} auch in Bezug auf die Verknüpfungen \wedge , \vee ein Schrägverband.

Man kann Funktionen mit den Eigenschaften (20) in mannigfacher Weise aufstellen, indem man spezielle Strukturen zugrunde legt. Benutzt man insbesondere geeignete Verbände, so erhält man Beispiele von Schrägverbänden aufgrund der Kenntnis von Verbänden.

Eine speziellere Klasse von Funktionen f , F erfüllt die oberste Zeile (20) in der Form

$$\left. \begin{aligned} f(a \cup b) &= fa \cup fb, \\ ffa &= fa. \end{aligned} \right\} \quad (21)$$

Entsprechendes ist für Fx zu sagen. Wenn \mathfrak{B} ein Verband ist, so ergibt sich bei dieser spezielleren Form (21) der Funktionen f , F übrigens *genau dann Erfüllung des Axioms (13), wenn*

$$Ffa \supseteq a; \quad fFa \subseteq a \quad (22)$$

ist.

Denken wir uns jetzt einen beliebigen Verband \mathfrak{B} mit Elementen a , b , ..., und bilden wir das direkte Produkt von \mathfrak{B} mit sich selbst, also einen Verband mit Elementen, welche *Paare* (a_1, a_2) von Elementen aus \mathfrak{B} sind. Daraus nehmen wir den Unterverband derjenigen Elemente, bei denen $a_1 \subseteq a_2$ ist. In dem so beschriebenen Verband \mathfrak{B} definieren wir zwecks Erfüllung von (21) und der entsprechenden Beziehungen für Fx :

$$\left. \begin{aligned} f(a_1, a_2) &= (a_1, a_1), \\ F(a_1, a_2) &= (a_2, a_2). \end{aligned} \right\} \quad (23)$$

Dieses ganz spezielle Beispiel einer Klasse von Schrägverbänden erfüllt übrigens auch das bemerkenswerte zusätzliche Axiom

$$\left. \begin{aligned} (a \wedge b) \vee (b \wedge a) &= (b \wedge a) \vee (a \wedge b), \\ (a \vee b) \wedge (b \vee a) &= (b \vee a) \wedge (a \vee b), \end{aligned} \right\} \quad (24)$$

welches passend als das Axiom *halbkommutativer* Schrägverbände bezeichnet werden kann, da es insbesondere immer dann erfüllt ist, wenn mindestens *eine* der beiden Verknüpfungen \wedge , \vee *kommutativ* ist.

Die mit der f , F -Konstruktion aus Verbänden abzuleitenden Halbverbände haben freilich trotz ihrer großen Mannigfaltigkeit eine ihnen gemeinsame sehr spezielle Eigenschaft: Sie erfüllen das zusätzliche Axiom (eine Verschärfung von (15)):

$$\left. \begin{aligned} a \wedge b \wedge c &= a \wedge c \wedge b, \\ a \vee b \vee c &= b \vee a \vee c. \end{aligned} \right\} \quad (25)$$

Ein Beispiel eines *modularen* Schrägverbandes, welcher dieses Zusatzaxiom (25) *nicht* erfüllt (wohl aber (15) erfüllt), ist durch folgende Verknüpfungstabelle für die vier Elemente 0, u , v , 1 des Schrägverbandes \mathfrak{B}_4 gegeben, in welcher x ein beliebiges Element von \mathfrak{B}_4 bezeichnet:

$$\left. \begin{array}{l|l} 0 \wedge x = 0 & x \vee 1 = 1 \\ u \wedge x = u & x \vee u = u \\ v \wedge x = v & x \vee v = v \\ 1 \wedge x = x & x \vee 0 = x \end{array} \right\} \quad (26)$$

Dieser Schrägverband \mathfrak{B}_4 ist für die Theorie der *distributiven* Schrägverbände von ähnlich grundsätzlicher Bedeutung, wie der aus nur zwei Elementen 0, 1 bestehende Verband für die Theorie der distributiven Verbände. Man kann allerdings den Begriff der distributiven Schrägverbände auf mannigfach verschiedene Weise definieren, derart, daß die Definition schärfer oder im Gegenteil toleranter gefaßt wird. Ein Beispiel eines *Distributivgesetzes* für Schrägverbände ist folgendes:

$$\left. \begin{aligned} c \wedge (b \vee a) &= c \wedge [b \vee (c \wedge a)], \\ [(a \vee c) \wedge b] \vee c &= (a \wedge b) \vee c. \end{aligned} \right\} \quad (27)$$

Dieses sehr tolerante Distributivgesetz — welches im kommutativen Fall

mit (7) gleichbedeutend wird — wird durch umfangreiche Klassen von Schrägverbänden erfüllt, insbesondere auch durch \mathfrak{B}_4 . Die oben nach der f, F -Konstruktion mit (21) konstruierten Beispiele erfüllen, wenn für \mathfrak{B} ein distributiver Verband genommen wird, ebenfalls (27).

Viel schärfere Distributivgesetze für Schrägverbände bekommt man jedoch aus (8). Wegen des kommutativen Gesetzes kann man (8) offenbar in 384 verschiedenen Formen schreiben, und vielleicht haben alle diese 384 verschiedenen Schreibweisen von (8) verschiedene Bedeutung, wenn sie mit Zeichen \wedge, \vee statt \cap, \cup geschrieben werden.

Aus Gründen, deren Erläuterung hier etwas zuviel Raum beanspruchen würde, kann man jedoch nur 6 von diesen 384 Formen als vermutlich bedeutungsvoll ansehen. Diese 6 Distributivgesetze sind nicht sämtlich gleichwertig; ob einige unter ihnen gleichwertig sein mögen, ist noch unentschieden. Der durch (26) definierte Schrägverband \mathfrak{B}_4 erfüllt alle 6 Beziehungen, und überdies noch 14 weitere, weil es in \mathfrak{B}_4 einige Übereinstimmungen gibt, die im kommutativen Fall trivial sind, aber im nichtkommutativen Fall keineswegs. Diese Beziehungen sollen unten zusammengefaßt werden.

Zuvor jedoch sei zur Erläuterung der besonderen Bedeutung von \mathfrak{B}_4 noch erwähnt: Bekanntlich kann jeder distributive Verband als Unterverband erhalten werden aus einem direkten Produkt, dessen Faktoren sämtlich dem aus zwei Elementen bestehenden Verband 0, 1 entsprechen. Analog kann eine weite, durch ein bestimmtes Konstruktionsverfahren definierte Klasse von Schrägverbänden erhalten werden durch Aussonderung von Unterbereichen aus solchen Schrägverbänden, welche als direkte Produkte von direkten Faktoren \mathfrak{B}_4 entstehen. Man kann deshalb *diese* Schrägverbände — die also alle im Folgenden verzeichneten Eigenschaften von \mathfrak{B}_4 ebenfalls besitzen — wohl als die im *schärfsten* Sinne „*distributiven*“ Schrägverbände bezeichnen.

Alle erwähnten Feststellungen — die nur einen kleinen Ausschnitt aus umfangreicheren Ergebnissen bilden — lassen uns freilich noch immer weit entfernt bleiben von dem mir vorschwebenden Ziel, welches angedeutet werden könnte als die *Konstruktion von verallgemeinerten projektiven Geometrien*, deren Elemente nicht mehr modulare Verbände, sondern modulare *Schrägverbände* bilden. Erst danach wird man beurteilen können, ob die Theorie der Schrägverbände, abgesehen davon, daß sie ein reizvolles Gebiet mathematischer Untersuchung zu ergeben scheint, auch für die Physik förderlich sein könnte.

Folgende Zusatz-Axiome I, II werden durch \mathfrak{B}_4 erfüllt:

I) *Folgende acht Polynome stimmen überein:*

$$\left. \begin{aligned} (b \wedge c) \vee (a \wedge b) \vee (a \wedge c) &= (b \vee a) \wedge (c \vee a) \wedge (b \vee c) \\ \Rightarrow (a \wedge b) \vee (b \wedge c) \vee (a \wedge c) &= (b \vee a) \wedge (b \vee c) \wedge (c \vee a) \\ \Rightarrow (b \wedge a) \vee (b \wedge c) \vee (a \wedge c) &= (b \vee a) \wedge (b \vee c) \wedge (a \vee c) \\ \Rightarrow (b \wedge c) \vee (b \wedge a) \vee (a \wedge c) &= (b \vee a) \wedge (a \vee c) \wedge (b \vee c). \end{aligned} \right\} \quad (28)$$

II) *Folgende vier Polynome stimmen überein:*

$$\left. \begin{aligned} (a \wedge b) \vee (c \wedge b) \vee (a \wedge c) &= (c \vee a) \wedge (b \vee c) \wedge (b \vee a) \\ \Rightarrow (c \wedge b) \vee (a \wedge b) \vee (a \wedge c) &= (c \vee a) \wedge (b \vee a) \wedge (b \vee c). \end{aligned} \right\} \quad (29)$$

Bibliographie

- [1] JORDAN, P., *Über eine nicht-desarguessche ebene projektive Geometrie*. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg, vol. 16 (1949), pp. 74–76.
- [2] ———, *Zur Theorie der Cayley-Größen*. Akademie der Wissenschaften und der Literatur. Abhandlungen der Mathematisch-Naturwissenschaftlichen Klasse. Series 3 (1949), pp.
- [3] ———, *Die Theorie der Schrägverbände*. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg, vol. 21 (1957), pp. 127–138.
- [4] ———, J. v. NEUMANN and E. WIGNER, *On an algebraic generalization of the quantum mechanical formalism*. Annals of Mathematics, vol. 35, (1934), pp. 29–64.
- [5] KOECHER, M., *Analysis in reellen Jordan-Algebren*. Nachrichten der Akademie der Wissenschaften in Göttingen, Series IIa, Nr. 4 (1958), pp. 67–74.

LOGICAL STRUCTURE OF PHYSICAL THEORIES

PAULETTE FÉVRIER

Henri Poincaré Institute, Paris, France

At the present time, the methodology of Theoretical Physics is not yet well determined and clear. There are various conceptions of it according to the different physicists, and, for some of them, the axiomatisation of the theories is only a part of the development of Theoretical Physics.

Adequacy is the fundamental notion from the theoretical physical point of view. A theory is adequate in a certain experimental domain if the predictions provided by this theory on the basis of given experimental data taken within this domain, agree with experiment. This reference to experiment introduces notions we do not meet about mathematical theories.

Let us consider a part of a physical theory which is axiomatised in a suitable way. Let us leave aside the physical meaning of the terms used in this theory. Then we get a certain mathematical theory. This theory possesses a *structure* (in Bourbaki's meaning); we shall call it the formal structure of the part of the physical theory under consideration.

If we have been able to axiomatise the whole theory, it receives a formal structure in the meaning above. But one should not forget that, in a physical theory, the terms *must have a physical meaning*, which is nothing else but an intuitive meaning. This meaning being left apart, one would have only a formal model left which would lose its interest for the physicist. That is why the meaning must always be taken into account together with the structure.

Many authors worked out more or less precise axiomatisations of wave mechanics or quantum theories, every one of which has its advantages and drawbacks, but an axiomatisation which would be in the same time completely satisfactory and adequate, seems not yet to have been proposed. I mean that, independently of the difficulties regarding the formal expression, which we shall leave aside as if they were resolved, an axiomatisation must not allow any example of inadequacy, i.e. a physical case which should be described by the axiomatised theory and yet escapes this description. One could give some examples of

particulars cases of inadequacy that have been mentioned about certain attempts at axiomatisation of waves mechanics:

a) certain axiomatic systems show a lack of adequacy because of such a potential V that the hamiltonian operator of the Schrödinger's equation does not possess any longer the general properties which are required of the operators associated with physical observables.

b) on the other hand it may be useful to examine whether some other axiomatic systems have not to be modified or improved because of the spectra which have points of accumulation (For instance, see Colmez' paper [3]).

In spite of these difficulties, it is possible to realise axiomatisations of some parts of a physical theory, but the main problem from a theoretical physical point of view is *to come to a better theory* rather than to a perfect axiomatisation of a given one. It is a matter of fact that a theory which would interest a physicist, is never completely built up, presents some defects, whereas a well-shaped theory in some way achieved does not attract him any longer, probably because, when this stage is reached, he already runs towards some other new growing theory. Processes of formation of new theories, that is the interest of the physicist, whereas the logician and the mathematician care for formal achievement.

Every physical theory holds in a limited experimental domain; the problem which is always before the physicist is to find new conceptions leading him to a new theory adequate to the experimental data unaccounted for by the preceeding ones.

Hence, if physico-logical studies can be useful for the physicist, they will be useful provided they are applied to *a theory not completely achieved* but still in the course of its development. When once the theory is quite built up, its inadequacies appear, the boundaries of its experimental domain are known, and the physicist turns himself towards the building of a new better theory. That is the reason why, from the special standpoint of the physicist, it may be more useful to elaborate psychological considerations in order to help his attempts at new theories, than to try to provide, for an achieved theory, a satisfactory axiomatisation in the most strict sense of the term.

However, the properly axiomatic enquiries about a given physical theory are necessary, not only from a formal point of view, but also from the point of view of the theory of knowledge.

Whatever point of view we adopt, it seems to me that the first difficulty which rises is to determine exactly what we mean by a physical theory and by a satisfactory axiomatisation of a physical theory. *What does the physicist intend when he tries to elaborate a physical theory?* This question appears as very important because, if we look at the considerations I mentioned before, we can find that they are not all related to the same meaning of the idea of a physical theory. It seems to me that such a question can be answered in three quite different ways:

1) the aim of the physicist when he makes a new theory can be only to find new results, that is to build up, at any rate and by any means, a theory which enables him to *predict some new experimental datum*;

2) the aim of a physical theory can be to provide what we call *an explanation of physical reality*, that is a formal construction, adequate to the experimental data, which connects them in a satisfactory rational way. Presently, what we should call a "rational way" of building explanations means a deductive way, *according to the axiomatic method*. This conception of a physical theory does not exclude the research for adequate predictions, but puts the formal requirements in first place;

3) the aim of a physical theory can also be *a description of physical reality* in the sense of a connexion between the set of experimental data, and some *principles and notions which are intuitively considered as fundamental* in something like a "Weltanschauung". These primitive elements must lead deductively to statements verified by experiment and they aim also to supply adequate predictions, but they are chosen first with respect to their fundamental role in the description. They rise from various previous considerations which form, according to Destouches' expression, an "inductive synthesis" a, p. 86; 5b, vol. I, p. 114.

Several axiomatic systems can be set up with respect to the same experimental domain; according to the second meaning among a physical theory, the best theories are the most suitable among these various axiomatic systems with respect to the requirements of the axiomatic method. According to the third meaning of a physical theory, the best theory is not necessarily the best system from an axiomatic point of view, but that, among the various systems, which depends on the most fundamental notions in an heuristic sense.

If we come back to the first of the three preceeding meanings, we see that it has to be considered as a *minimum requirement* with respect to the question: what is a physical theory? Indeed, it founds a physical

theory on the single purpose of calculating predictions for future experimental data, starting from initial experimental data.

Taking this minimum requirement as a primitive assumption, one can build up the most general physical theory, that is a *general theory of predictions* [5b, vol. II, pp. 505–654, vol. III, 705–742; 5c]. A summarization of this theory has been given by Destouches in his lecture. It aims to be a frame in which will enter any physical theory, I mean it aims to point out what, in any physical theory, is involved in the particular initial purpose of calculating predictions.

This general theory of predictions is, by definition, a physical theory in the first sense of the term, but, according to the second one, it can be axiomatised. According to the third meaning, the general theory of predictions points to the purpose of calculating predictions as one of the most fundamental notions in a physical theory, if we consider a physical theory as an attempt made by a physicist in order to provide a “Weltanschauung” adequate, not only to the experimental data already known, but also to future experimental data. However, a physicist who assumes the third meaning of a physical theory requires more than the single idea of adequate prediction to set up a particular physical theory.

The physico-logical studies do not restrict themselves to one of these three points of view. That is why they are not formal on the whole, though many parts of them can be formalised. They do not pretend to be more than a help, as well for the approaches of the physicist as for those of the metamathematician or of the philosopher.

The way in which physico-logical studies may contribute to elaborate physical theories is the following: in order to satisfy a certain physical condition by means of a theory that we try to elaborate, physico-logical considerations can supply the theoretical requirements to be fulfilled by this theory. When the physical theory must satisfy several physical conditions, some of them being in contradiction, physico-logical considerations permit us to reduce the contradictions, and to establish which elements of the theory remain to be determined, in order to achieve it.

I shall try now to give some examples of physico-logical enquiries, in the sense explained above.

The first task to set up a physical theory is to elaborate schemes of the concrete physical operations as: making a measurement, reading the result of a measurement, etc. . . . For example, from a schematic point of

view, a measurement which is supposed but not effectively realised can be assimilated to an affective measurement; we can also assimilate a result of measurement to a prediction for the very instant when the measurement is effected.

Further, we have to represent such schemes by suitable mathematical entities, as sets, elements, etc. In that way, let us consider more thoroughly the initial assumption on which the general theory of predictions is based. We have to make precise what we mean by an "initial datum" and by a "prediction statement", and, more generally, to determine what are the various kinds of statements which have to be taken into account in a physical theory.

Measurements are classified into types called *observables*, and determined by various experimental processes. One observable is represented by an element of a set called *set of the observables*. An experimental datum is read by the observer on the dial or the scale of a measuring apparatus (for example, it is the position of the spot in a galvanometer). This position is not infinitely precise. In a schematic way, we can represent it by an interval E with rational ends on a straight line or a circle. Its extent is appreciated by the observer on the basis of all that he knows about the precision of his apparatus. For instance, the precision is obtained by repeating a particular measurement a rather large number of times (we know that its results might be always the same), and by calculating the standard deviation of the various numbers obtained in this way. The weakest assumption we can make about it is to assume that the result of the measurement is this very interval E and not a special

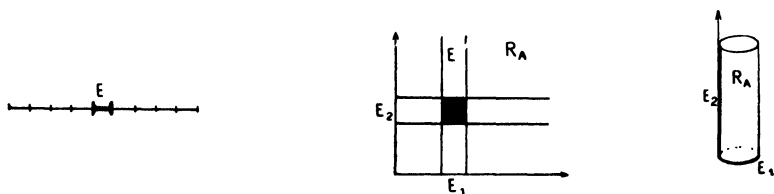


Fig. 1

point, determined but unknown, inside that interval. In the case of macroscopic theories we can admit that there is in E one point which is the real result of the measurement, but in microphysics such an assumption cannot be made.

Instead of an apparatus with only one dial we may have an apparatus with several dials, which can be linear or circular. For every reading of a dial we shall have an interval on a straight line or on a circle. In order to get the complete result of the measurement, which is an n -interval with rational ends, all the dials must be taken into account at the same time, and then the result is represented in the cartesian product of the curves on which have been represented the results taken from every dial. This cartesian product makes up a space called "*observational space of the observable A* ", denoted by (R_A) , and such a space is associated with each observable [1] (see figure 1).

A sentence which states an experimental datum is an *empirical sentence* such as

$$\text{at } t_0, \text{ Re Mes } A \subseteq E_A$$

where t_0 is an instant of the observer's clock, $\text{Re Mes } A$ is a certain set in (R_A) , and E_A a specified set in (R_A) .

Though not any specified theory is assumed when we begin to set up a general theory of predictions, however we cannot give a physical meaning to an empirical sentence without admitting that such a meaning is provided to this empirical sentence by a certain theory, the theory by means of which the experiment has been motivated.

In order to take into account the case of a theory with a quantization, we have to introduce the set \mathcal{A}_0 of the possible values of an observable A , which is a set in the observational space (R_A) . In the particular case of no quantization, as in macroscopic physics,

$$\mathcal{A} = R_A$$

Hence, from an empirical sentence, we obtain a so-called *experimental sentence* by intersection of E and \mathcal{A} , that is

$$\text{at } t_0, \text{ Re Mes } A \subseteq \mathcal{E} \text{ (where } \mathcal{E} = E \cap \mathcal{A}\text{)}.$$

In the case of no quantization

$$\mathcal{E} = E.$$

As I have already said, we may in theoretical physics take under consideration not only statements expressing facts effectively realized, but also statements concerning supposed facts. From a schematic point of view we can look at these supposed data in the same way as the real data.

Then, from an experimental sentence or a pair of experimental sentences about one observable A , we can yield by logical means new sentences which will be also experimental sentences. In this way, we obtain a *calculus for the experimental sentences* concerning only one observable A . We can then look at this calculus as a formal system. For example, we can denote by

p_1 the sentence: at t_0 , $\text{Re Mes } A \subseteq \mathcal{E}_{p_1}$

p_2 the sentence: at t_0 , $\text{Re Mes } A \subseteq \mathcal{E}_{p_2}$

and define

$p_1 \& p_2 =_d \text{Re Mes } A \subseteq (\mathcal{E}_{p_1} \cap \mathcal{E}_{p_2})$ (logical product)

$p_1 \vee p_2 =_d \text{Re Mes } A \subseteq (\mathcal{E}_{p_1} \cup \mathcal{E}_{p_2})$ (sum or superposition)

(See figure 2)

$\sim p =_d \text{Re Mes } A \subseteq (\mathcal{A} - \mathcal{E}_p)$ (negation)

From these definitions arise the rules of the calculus of experimental sentences for one observable. When it is formalized, this system is a

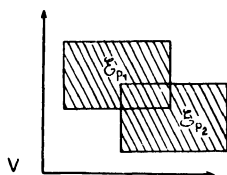


Fig. 2

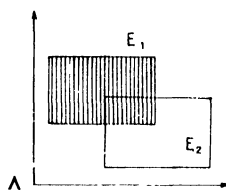


Fig. 3

language $L_{1,A}$ of the experimental sentences. It is obviously a *boolean algebra* [7a].

Now, another case must be examined. Because of a lack of precision when we read the result of the measurement, it is possible that we cannot state if it is the n -interval E_1 or another n -interval E_2 which is the result really indicated by the dial, and we can state only that the result is: E_1 or E_2 . In that case, the datum is no longer expressed by an experimental sentence, since the result is no more one set, but one set *or* another set. However, we know something about this result, which can be expressed by a kind of proposition of a more general type than the experimental sentence defined above, and that we shall call an *experimental propositional*, according to a suggestion of Prof. Beth (see figure 3).

The propositional expressing that the result of the measurement is either E_1 or E_2 will be denoted by $p_1 \vee p_2$, and every propositional can be obtained from experimental sentences by means of that operation \vee called *mixture or strong logical sum*.

Experimental sentences may be considered as particular propositionals. When we formalize the system obtained in that way we have a language $L_{2,A}$ of the propositionals for the observable A .

Now we can take experimental sentences concerning several observables A, B, \dots . If these observables can be observed at the same time, we form a compound observable that I shall denote $A \S B$, and we can reduce to the case above of one observable; \S is a binary operation which is applied to certain pairs of the set of the observables. If there is a single pair of observables which cannot be measured together (as in micro-physics), we have to bring in the calculation of predictions, in order to be able to describe that special case.

As Destouches explained in his lecture, the general theory of predictions enables us to point out a correspondence between every experimental sentence and a subspace \mathcal{M}_p passing through the origin of a vector space (\mathcal{V}). If that space (\mathcal{V}) would have a finite number of dimensions, then all observables would have a finite spectrum; hence, adequacy requires that the space (\mathcal{V}) be infinite dimensional. When an operation applied to experimental sentences yields an experimental sentence p , a subspace \mathcal{M}_p corresponds to this sentence and the operation induces in the space (\mathcal{V}) an operation on the subspaces.

Then the properties of the sentential calculus on the experimental sentences will be those of the calculus on the associated subspaces. The study of these properties enables us to point out the characteristics of the theory elaborated in order to calculate predictions for future measurements. Two cases are to be considered:

1) in the case of one observable, or of observables each pair of which can be measured at the same time, the set of the associated subspaces is a Boolean algebra. Hence, in a physical theory where all observables can be measured at the same time, the experimental sentences follow the rules of the *classical sentential calculus*;

2) in the second case, there is at least one pair of observables which, by right, cannot be measured at the same time. "By right" means: according to the theory. In that case, the study of the operations on the subspaces associated with the experimental sentences points out the

characteristics of the corresponding logical operations, in such a way that the logic which is then adequate is no longer the classical sentential calculus, but a special logic L_{CS} with the following rules for the logical negation, product and sum [1; 7a, pp. 91–216]:

NEGATION: To the negation $\neg p$ of p corresponds the sentence asserting that the result of the measurement is not in \mathcal{E}_A ; hence the associated subspace is the complementary orthogonal subspace.

LOGICAL PRODUCT: Because of the homomorphism between the sentential calculus and the calculus on the subspaces \mathcal{M} , with a sentence r which is the logical product of p and q

$$p \& q = r,$$

is associated the intersection of the corresponding subspaces

$$\mathcal{M}_{p\&q} = \mathcal{M}_p \cap \mathcal{M}_q$$

But, in the case where p and q relate to observables which, by right, cannot be measured at the same time, the corresponding subspace $\mathcal{M}_{p\&q}$ is reduced to the point 0. Hence the sentences of type $p \& q$ are *excluded*, that is certain pairs of sentences are *not compossible*.

LOGICAL SUM: The conjunction “or” joining two experimental sentences can take two different meanings:

Superposition: We can define a *weak logical sum* $p \vee q$ with the following meaning: a measurement of A has been effected, or a measurement of B , or a measurement of the compound observable $A \& B$, with such an imprecision that we can only assert that the result of the measurement belongs to

$$(\mathcal{E}_A \times \mathcal{A}_B) \cup (\mathcal{A}_A \times \mathcal{E}_B)$$

\mathcal{A}_A and \mathcal{A}_B being the spectra of A and B . To this operation corresponds the *sum of subspaces*

$$\mathcal{M}_{p\vee q} = \mathcal{M}_p \oplus \mathcal{M}_q$$

$\mathcal{M}_{p\vee q}$ being the subspace spanned by \mathcal{M}_p and \mathcal{M}_q in (\mathcal{Y}) .

In wave mechanics, this operation describes the notion of *superposition*.

In this way, the logical operations $\&$, \vee , with \neg , have the same properties as the operations \cap , \oplus and orthocomplementation of the subspaces passing through 0 of the space (\mathcal{Y}) . Thus the sentential calculus appears as isomorphic to an *algebra of infinite dimensional ortho-complemented*

projective geometry. It is an *ortho-complemented lattice*, non modular in the general case.

Mixture: On the other hand, the *mixture* of experimental sentences leads us to a calculus of experimental propositionals, in the following way:

With the sentences p, q , we can associate a sentence $p \vee q$ called *strong logical sum*. It means that either the observable A has been measured, and the result has been found in \mathcal{E}_A , or the observable B has been measured, and the result has been found in \mathcal{E}_B ; or both observables have been measured if they are compossible, and the result has been found in $\mathcal{E}_A \times \mathcal{E}_B$, but we do not know which one of these three cases has been realized. To this strong logical sum \vee corresponds the *union* of the associated subspaces, which is not a subspace:

$$\mathcal{M}_{p \vee q} = \mathcal{M}_p \cup \mathcal{M}_q.$$

Hence we are lead to distinguish from the sentential calculus a *calculus of propositionals*, a *propositional* being a strong logical sum of sentences. This is the *language* L_4 of the experimental propositionals, which is a distributive lattice in \wedge, \vee .

Now, we may have to express that, for instance, if p is asserted, then q has also to be asserted. That introduces a relation \rightarrow in a *language* L_5 . To the relation \rightarrow corresponds the relation of *inclusion* for the corresponding subspaces. In the same way, we shall have a *language* L_6 for the propositionals; and in that case, to the symbol \rightarrow will correspond the *inclusion of the sets formed by union of subspaces*.

At last, we have a *language* L_7 which is the language of the physical theory under consideration. In order to understand what is the language L_7 , let us take the case of Newtonian mechanics: L_7 denotes what would be the formalization of Newtonian mechanics when the initial conditions are left free. (Here the experimental sentences state that the initial conditions belong to a certain set of the phase-space). In any physical theory, L_7 corresponds to the formalization of its deductive part.

We see that the general theory of predictions enables us to make appearant the logical structure of physical theories, by means of correspondence between certain subspaces and the various sets of sentences used in these theories. Moreover, the general theory of predictions shows thus that it must be distinguished between two kinds of physical theories, as Destouches says in his lecture. The calculus of experimental sentences of the quantum theories is not a boolean algebra but an algebra of pro-

jective geometry, and that is, in my opinion, the most important characteristic of the structure of this kind of theories.

I should like now to give another example of physico-logical considerations about the comparison between these two kinds of theories.

An historical exemple of this duality of physical theories is given now by the opposition between the so called classical probabilistic interpretation of wave mechanics, and the causal and deterministic interpretation proposed, some years ago, by David Bohm [2], Louis de Broglie [4] and several other physicists.

I think that, from a physico-logical point of view, we do not have to decide in favour of one or of the other kind of theory, because it is possible, as I shall try to show now, to find means of translating one into the other and conversely [7b; 7c].

First, one can prove that *it is possible to pass from a quantum phenomenalist theory to a causal theory*, provided a modification be made of the notion of the physical system described.

Let S be a source of particles, for example an electrons-gun; in wave mechanics in its usual meaning, the system S that the theory plans to describe is one electron; in a causal theory, the observed system is determined only if we decide which experimental apparatus we put after the gun; for instance we can put a screen with one hole of a given diameter; this measuring apparatus a allows us to know, with a precision determined by the diameter of the hole, the value of the observable A which is the position of the particle; the system in observation is then S/a_A . We might put, instead of a screen with a single hole, a screen with two holes (Young's holes). Then we should have a quite different system, S/a_B , which cannot be realised at the same time as S/a_A . Thus, in a causal description, in these two cases we have two different physical systems; indeed, the boundary conditions are different, the quantum potentials are different. In both cases the parameters which can be reached by experiment are not the same. The initial conditions, in both descriptions, are the same: they are determined by the characteristics of the gun.

In the case of the usual quantum theory, the studied system is, as we have seen, the particle S ; what we know about the gun determines the initial wave; if we use the compound system-apparatus S/a_A , we measure the observable A on S ; if we use the compound system-apparatus S/a_B , we take a measurement on the observable B on S . Thus, the observed system is always the same system S .

The rules of wave mechanics supply predictions by means of probabilities concerning the results which will be obtained. As we have seen before, the set of the experimental sentences is a non-distributive, and generally non-modular, lattice. But it admits boolean sub-lattices B_A for every observable A . And *such a sub-lattice is identical to the boolean sub-lattice of the experimental sentences concerning the system S/α_A in the causal description*. This identity is what makes possible the duality of description but *the correspondence between the sentences of the two types of theories cannot be extended further than the case of the experimental sentences for a complete observable in the quantum description*. Indeed, the lattice of the experimental sentences of the probabilistic description is not distributive, while the algebra of the experimental sentences in the causal description is distributive. The correspondence can be extended only by means of probabilities: to an experimental sentence expressing a maximum observation on the system S (hence determining a single initial function) corresponds a law of probability for the observable A , hence a valuation of the boolean algebra B_A , and a law of repartition for the system S/α_A .

We see then that position plays a special role in a causal theory; indeed, A may be the position; if A is not the position, we shall observe the system $(S + \alpha_A)/\alpha_B$ where B is an observable reducible to a measurement of position; thus, by changing the physical system in observation, one can always reduce to position in a causal theory.

In this way, we see that the two kinds of theories are equivalent with respect to a certain experimental domain, or set of facts. We can pass from one to the other *providing a modification is made of our conception of the physical system taken under consideration*.

Conversely, it can be shown that a translation is possible from a causal theory to a probabilistic one. Let us assume a causal theory which supplies descriptions for the systems S/α_A , S/α_B , etc. Is it possible to build up a probabilistic theory supplying the same predictions as this causal theory but which would not contain parameters that we cannot reach by experiment. In such a theory, we have to take into account only the sentences corresponding to parameters which can be submitted to experiment. Hence the construction of the theory must be effected in two steps: 1) to select, among the initial sentences in the causal theory, and for every system S/α_A , S/α_B , etc (S remaining the same), the experimental sentences and their consequences. That can be made by a logical process using the modality "*experimentable*". This process enables us to show that the experimental sentences of the causal theory form a sub-set of the

lattice of the experimental sentences of the probabilistic theory.

2) Then we have to join in a single description concerning S the partial descriptions corresponding to every system S/a_A , etc. That can be realised; and then it is sufficient to identify the expressions of the probabilities computed according to the causal theory, and those computed according to the general theory of predictions. Since every theory supplying predictions can be put in the frame of the general theory of predictions, such an identification is possible.

Thus, if one has been able to set up an adequate causal theory in microphysics, one can, *by eliminating the elements of the theory which cannot be experimented upon*, build up a probabilistic theory, equivalent to the given theory in the following way: it supplies the same predictions about future measurements. Such a theory has the same structure as the usual quantum theory, and is essentially indeterministic. It does not contain non-experimentable observables.

From these two processes of translating one kind of theory into the other, we can see that they are not different with respect to experimental data or adequacy. Their difference, in fact, concerns methodological assumptions. *If one prefers a positivistic approach to elaborate physical theories, then one cannot admit in a theory physical entities which cannot be experimented upon, but the price of this is indeterminism. On the other hand, if one cannot accept indeterminism, one has to assume that certain physical entities escape experimentation.*

Bibliography

- [1] BIRKHOFF, G. and VON NEUMANN, J., *The logic of quantum mechanics*. Annals of Mathematics, vol. 37 (1936), pp. 823–843.
- [2] BOHM, D., *Suggested interpretation of the quantum theory in terms of "hidden" variables*. Physical Review, vol. 85 (1952), pp. 166–193; vol. 87 (1952), p. 389; vol. 89 (1953), p. 458.
- [3] COLMEZ, J., *Définition de l'opérateur H de Schrödinger pour l'atome d'hydrogène*. Annales scientifiques de l'Ecole Normale Supérieure, 3ème Série, vol. 72 (1955), pp. 111–149.
- [4] DE BROGLIE, LOUIS. a) *Sur la possibilité d'une interprétation causale et objective de la mécanique ondulatoire*. Comptes-rendus Acad. Sciences Paris, vol. 234 (1952), p. 265.
 · b) *La physique quantique restera-t-elle indéterministe?* Paris, 1953, VII + 113 pp.

- [5] DESTOUCHES, J. L., a) *Essai sur la forme générale des théories physiques*. Thèse principale pour le Doctorat ès Lettres, Paris, 1938; Monographies mathématiques de l'Université de Cluj (Roumanie), fasc. VII (1938).
b) *Principes fondamentaux de Physique théorique*. Paris, 1942, IV + 905 pp.
c) *Corpuscules et systèmes de corpuscules. Notions fondamentales*. Paris, 1941, 342 pp.
- [7] FEVRIER, P., a) *La structure des théories physiques*. Paris, 1951, XII + 424 pp.
b) *Sur l'élimination des paramètres cachés dans une théorie physique*. *Journal de Physique et Radium*, Vol. 14 (1953), p. 640.
c) *L'interprétation physique de la mécanique ondulatoire et des théories quantiques*. Paris, 1956, VIII + 216 pp.

PHYSICO-LOGICAL PROBLEMS

J. L. DESTOUCHES

Henri Poincaré Institute, Paris, France

1. **Introduction.** I call physico-logical problems not the purely logical ones, but those in which both logical conditions and some physical interpretation arise. About a physical theory there are various questions of this kind; but these questions are not yet studied in details and we have still to detect and specify the problems occurring and to build up suitable methods. I shall try to set up a general survey of physico-logical problems and to summarize the general theory of predictions.

2. **Formal considerations.** Let us take a physical theory which is considered as complete by a physicist. We can, like in the case of euclidean geometry, axiomatize and formalize it, and make about it the same formal enquiries as about a mathematical theory. However, when we consider a modern physical theory, it is in fact very difficult to elaborate a suitable axiomatic system. Very often an axiomatic system for a physical theory does not cover all physical cases; some exceptional case appears which does not enter the axiomatic scheme. Here I shall put aside this purely formal point of view, and consider only physico-logical problems.

3. **The three parts of a theory.** First of all, many people believe that a physical theory taken as a whole is a deductive theory, that is a theory based upon a few primitive terms and postulates and then developed in a strictly deductive way. But, in fact, things are not so easy: we find a mixture of physical notions which have to be clarified by degrees; and the physical theory will keep the imprint of the efforts which led to its formation. I have called this first stage the *inductive synthesis* of the theory [4], which bring us to the *axiomatic part* of the theory, itself the second stage. Then comes the *deductive stage*, the third one. But in fact, the preceeding description is still too easy. The primitive terms and the postulates are not introduced all together but progressively. The three stages are mixed up with-one-another. What is the deductive part of a subtheory is at the

same time a piece of the inductive synthesis of a more fully developed part of the whole theory.

Formalisation can only be applied to the deductive side of the theory; in particular, the whole inductive synthesis cannot be formalised, but only some parts of it.

4. Adequacy. In a physical theory, we cannot lose sight of the physical meaning of the terms; we shall therefore remain at the level of intuitive semantics; the requirement of *adequacy* to experiment dominates any study about the notion of a physical theory. Adequacy consists in the fact that the predictions calculated according to the considered theory, are not at variance with experiment. At best a theory is adequate in a certain field called the *adequacy-domain* of the theory [10; 4c, pp. 40–69].

5. Search for a new theory. The search for a better theory belongs to the normal development of theoretical physics. Physico-logical considerations allow us to find out whether a new theory should replace an older one; and to shape a theory better than given theories.

Processes of unification of given theories can be pointed out, whether these theories show mutual contradiction or not [5; 4b, pp. 122–147]. When we elaborate a physical theory, we generally have to take into account incompatible conditions. Various formal processes can be used to avoid the contradictions, but the difficulty lies in finding a formal process appropriate to the physical requirements.

6. Formal structure. To each physical theory (as well as to each part of a physical theory) corresponds a *formal structure* [3]: the structure of the formal mathematical system in which the theory is formulated. I shall call this formal mathematical system the *algorithm* of the theory. When we pass over to a better theory, or to the unification of several theories, a part of the formal structure of the preceeding theory is maintained [1]; it helps us to set up the new theory. For instance, if the law of connexions between observers remains the same when we pass from a theory Th_0 to a theory Th_1 , then the geometrical algorithm remains unchanged in the new theory [1; 7]. For example, in classical mechanics the geometrical algorithm is the vector-calculus in the field of real numbers, and in wave mechanics we have as the geometrical algorithm a weaker algorithm. It is necessarily a vector-calculus. On the other hand the general theory of predictions implies that to each observable corresponds

a linear operator. So this weaker algorithm is a vector calculus on a ring of operators.

Quite a large part of wave mechanics can be obtained by this process.

7. General theory of predictions. A more concrete level of the studies on physical theories appears when one takes into account the fact that the aims of a physical theory are, at the minimum, to calculate predictions about the results of future measurements, starting from the results of initial measurements. In that way, we are led to a *general theory of predictions* which has a great deal of consequences [6; 11; 10b, pp. 91–318]. If an initial experimental datum obtained by an observer Ob about an observable A on the physical system S at an instant t_0 on his clock is described by a set \mathcal{E}_A of the observational space (R_A) , $\mathcal{E}_A \subseteq (R_A)$; and if we are trying to calculate some prediction for the result of a measurement which will be realised at an instant t' by an observer Ob' (in the future) on the system S , this prediction will be expressed in terms of a function \mathfrak{P} , the arguments of which are of two kinds: 1°) what we know: A , \mathcal{E}_A , t_0 , and 2°) what we predict: the result of the measurement which can be obtained at the instant t' of the clock of Ob' by this observer Ob' and described by a set \mathcal{E}_B of the observational space (R_B) of the observable B , that is

$$(1) \quad \text{Prob}\{\text{RéMes } B \subseteq \mathcal{E}_B \text{ at } t' \text{ by } Ob' / \text{RéMes } A \subseteq \mathcal{E}_A \text{ at } t_0 \text{ by } Ob\} = \\ = \mathfrak{P}(A, \mathcal{E}_A, t_0, Ob; B, \mathcal{E}_B, t', Ob'; S)$$

The problem of prediction is the problem of the computation of the \mathfrak{P} -functions.

In the most simple case we have only one initial measurement and we consider only one observer (thus Ob' is the same observer as Ob). Here we limit ourselves to this case.

8. Axiomatisation of measurement. That is the intuitive formulation of the problem of prediction. We shall now describe this problem in a more precise and more formal way. The physical system shall be described by a constant S and a measurement by the predicate Mes ; a measurement on the system S at time t_0 with an apparatus α shall be described by

$$\text{Mes}(\alpha, S, t_0)$$

this, being a primitive term. We admit now:

POSTULATE 1: *To each apparatus α corresponds an element A of a set T called "observable" or "type of measurement".*

POSTULATE 2: *To each measurement at time t_0 of type A corresponds a set \mathcal{E}_A which is a subset of \mathcal{A}_{A, t_0} called "spectrum of A at t_0 " and $\mathcal{E}_A = E_1 \times E_2 \times E_3 \times \dots \times E_n$. The E_i are rational intervals of finite sets or enumerable sets.*

In this case we write

$$\text{RéMes}(\alpha_A, S, t_0) \subseteq \mathcal{E}_A$$

and this is called an *experimental sentence*. So α_{A, t_0} called *spectrum* of A , is a subset of an n -dimensional space (R_A) called the *observational space* of A .

POSTULATE 3: *The number n of the sets \mathcal{E}_i depends only on the type of measurement: $n = \varphi(A)$.*

9. Axiomatisation of prediction. For an observable B , we consider the field E of the probabilisable (or measurable) subsets of the spectrum \mathcal{A}_B of B . We call a *probability* for $\text{RéMes}(\alpha_B, S, t) \subseteq \mathcal{E}_B$ where $\mathcal{E}_B \in E$ the value of a function $\mathfrak{P}(\mathcal{E}_B)$ defined on E and such that

$$1^\circ) \quad 0 \leq \mathfrak{P}(\mathcal{E}_B) \leq 1$$

$$2^\circ) \quad \mathfrak{P}(\mathcal{A}_B) = 1$$

$$3^\circ) \quad \mathfrak{P} \text{ is completely additive: } \mathfrak{P}(\Sigma \mathcal{E}_i) = \Sigma \mathfrak{P}(\mathcal{E}_i) \text{ if } \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \text{ for all } ij$$

4°) \mathfrak{P} depends on the measured observable A , the result \mathcal{E}_A of the measurement, the time t_0 of this measurement, the observable B , the set \mathcal{E}_B , the time t when this measurement shall be made, and the system S .

POSTULATE 4: *For a system S there exists at least one function \mathfrak{P} satisfying the conditions 1° – 4° .*

For each physical theory there is a set of \mathfrak{P} -functions which fulfill the conditions 1° – 4° under the conditions fixed by the principles of this theory. Conversely, any supplementary condition on the \mathfrak{P} -functions defines a class of physical theories. Thus we have a frame to discuss general properties of a physical theory.

10. Initial elements and prediction elements. To calculate the suitable \mathfrak{P} -functions, I proved [6; 11; 10b, pp. 91–318] that it is possible to do as follows:

First the initial experimental data are translated into an abstract language in which a set $\mathcal{X}_{0,A,\mathcal{E}_A,t_0}$ of abstract elements called "initial elements" corresponds to the datum

$$(2) \quad \mathcal{X}_{0,A,\mathcal{E}_A,t_0} = \Phi(A, \mathcal{E}_A, t_0, S) \text{ and } \mathcal{X}_{0,A,\mathcal{E}_A,t_0} \subseteq \mathcal{X}_0$$

(In wave mechanics, the set $\mathcal{X}_{0,A,\mathcal{E}_A,t_0}$ reduces to the set of the initial wave-functions $\{\psi_0\}$ compatible with the result \mathcal{E}_A , and \mathcal{X}_0 is the sphere of unit radius in a Hilbert space). Then an initial element X_0 belonging to $\mathcal{X}_{0,A,\mathcal{E}_A,t_0}$

$$X_0 \in \mathcal{X}_{0,A,\mathcal{E}_A,t_0}$$

is transformed at the instant t into another abstract element $X(t)$ called the "prediction element" by a one-one transformation $\mathfrak{U}(t, t_0)$ such that

$$(3) \quad X(t) = \mathfrak{U}(t, t_0)X_0$$

(In wave mechanics $X(t)$ reduces to a wave function $\psi(t)$).

Then the probabilities for the result \mathcal{E}_B for a future measurement can be calculated by a time-independent function F :

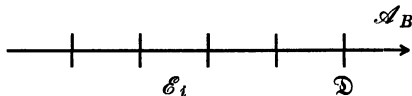
$$(4) \quad \mathfrak{P}(A, \mathcal{E}_A, t_0, Ob; B, \mathcal{E}_B, t, Ob; S) = F(B, \mathcal{E}_B, X; S, Ob).$$

Formulas (2) and (3) are the result of the use of the auxiliary-variables-method; (4) is a condition imposed on the evolution operator \mathfrak{U} .

The set \mathcal{X} of all X -elements can be considered as a subset of an abstract vector-space (\mathcal{V}) .

If that space (\mathcal{V}) would have a finite number of dimensions, then all observables would have a finite spectrum. Hence, adequacy to experiment requires that the space (\mathcal{V}) be infinite dimensional.

11. Decomposition of a spectrum. If \mathfrak{D} is a decomposition of the B-spectrum and \mathcal{E}_t an element of \mathfrak{D}



there exists at least one X_t in \mathcal{X} for which

$$(5) \quad F(B, \mathcal{E}_t, X_t; S, Ob) = 1$$

that is an X_t which guarantees that the B-experimental datum shall be

included in \mathcal{E}_i . These X_i can be defined as eigenfunctions of a linear operator in (\mathcal{V}) . Therefore an operator is associated with each observable by a formal process, (without any physical hypothesis; the physical content is introduced by the analytical form of the operator when this form is given explicitly) [6; 11; 10b, pp. 91–318].

It is possible to define an equivalence modulo B , \mathfrak{D} in which

$$(6) \quad X \equiv \sum_i c_i X_i \bmod B, \mathfrak{D}$$

In many cases, with a convenient definition of an abstract integral [6h; 18; 19], when there is a continuous part in the spectrum for B , the sum $\sum_i c_i X_i$ has a limit when we consider the set $\Pi\mathfrak{D}$ of every decomposition of the B -spectrum; in this case we have

$$(7) \quad X \equiv \int_{\mathcal{A}_B} c(d\mathcal{E}) X(d\mathcal{E}), \bmod B, \Pi\mathfrak{D}$$

12. The spectral decomposition theorem. In (6), c_i is a complex number and there exists at least one function $f_B \mathfrak{D}$ for which

$$(8) \quad F(B, \mathcal{E}_i, X; S, Ob) = f_B \mathfrak{D}(c_i)$$

It is possible to choose a function f_B independant of the decomposition \mathfrak{D} of the B -spectrum [6i, pp. 529–538]; in this case the function f_B must be a solution of the Cauchy's equation

$$f_B(x) \cdot f_B(y) = f_B(xy)$$

and fulfills some accessory conditions like $f(0) = 0$. If we exclude the total discontinuous solutions of Hamel, the only (continuous) solutions are

$$f_B(c_i) = |c_i|^k \text{ and } k > 0.$$

Moreover there exists one and only one universal function f independent of B and \mathfrak{D} when there is a pair of observables which are not simultaneously measurable [6i, pp. 538–540; 10b, pp. 221–233], that is a unique value for the constant k which is the same for all observables B . In the case where all observable are simultaneously measurable (as in classical physics) the value of k remains undetermined under the condition $k > 0$.

The physical consequence of this fact is that in classical physics, there exist no interferences of probabilities; on the contrary in quantum

physics, where non-simultaneously measurable observables exist, there are interferences of probabilities. Hence the value of k is an important property of a physical theory with non-simultaneously measurable observables.

P. Février has proved [12] that the constant k is equal to 2, so that the following spectral-decomposition theorem is valid:

THEOREM: *In the case where there exists at least one pair of non simultaneously measurable observables, the universal function f is*

$$f(c_i) = |c_i|^2;$$

hence $k = 2$.

In this case where there exists a non-simultaneous pair of observables, it can be proved that the general formalism of predictions cannot be reduced to a simpler one. On the contrary, when all observables are simultaneously measurable, (in this case the value of k remains arbitrary under the condition $k > 0$, and in particular we can put $k = 2$), the general formalism of prediction calculus is valid, but it can be reduced to a simpler one, that is a phase-space scheme.

So there are two types of physical theories, and only two: in the first type, there is at least one non-simultaneously measurable pair of observables; in the second type, all observables are simultaneously measurable. The classical theories are of the last type, and the quantum ones are of the first type.

13. Miscellaneous notions. 1°) A theory is called *objectivistic* if it is possible to eliminate apparatus of measurement from the theoretical formulation of phenomena. In this case the formalism of prediction calculus is reducible to a phase-space-scheme, and thus is of the second type.

On the contrary, a theory is called *subjectivistic* if an essential role is played by observers and apparatus of measurement in the theoretical formulation of the phenomena. This intuitive definition is interpreted formally as "the general formalism of prediction calculus for this theory is not reducible". It results from the above that a subjectivistic theory is of the first type; reciprocally a theory of the first type is subjectivistic.

2°) An observable B *derives* from an observable A if it is possible to compute the value of B at t_0 when the result of a measurement of A' at t_0 is known. A theory admits a *state-observable* if there exists an observable

such that all observables derive from it. In the other case, a theory is *without state-observables*.

It can be proved that if a theory is without state-observables this theory has at least one pair of non simultaneously measurable observables and so is of the first type. It is obvious that a theory with a state-observable has all observables simultaneously measurable and is of the second type.

3°) An experimental datum is a result of a measurement; it depends on the observed system and on the apparatus of measurement. Then if an experimental datum is an *intrinsic property* of the observed system, this experimental datum is independent of the apparatus of measurement. If all experimental data are intrinsic properties of the observed physical system, then the apparatus of measurement does not play an essential role and the theory is objectivistic.

On the contrary, if the experimental data are not intrinsic properties of the system, they depend on the apparatus and they play an essential role in the theoretical description, so that the theory is subjectivistic.

4°) An imprecise experimental datum is *analysable*, if it is equivalent either to consider the result \mathcal{E} of the measurement (\mathcal{E} is a set, see postulate 2), or to consider the result \mathcal{E}_1 or the result \mathcal{E}_2 , when $\mathcal{E}_1 \cup \mathcal{E}_2 = \mathcal{E}$. In other terms, a result of a measurement is analysable if for every prediction it is equivalent to consider the experimental sentence p corresponding to \mathcal{E} , or to consider the logical sum $p_1 \vee p_2$ (where p_1 corresponds to \mathcal{E}_1 and p_2 to \mathcal{E}_2).

When an imprecise experimental datum is not analysable, it is impossible to attribute a precise but unknown value to the measured observable. By means of the connexion between experimental sentences and closed linear manifolds in the space (\mathcal{U}) it can be proved [10b, pp. 156–159, 275–280] that, if the imprecise experimental data are all analysable, then the theory is objectivistic, and if there is some non-analysable experimental datum, then the theory is subjectivistic.

5°) The term *by right* means: “with respect to the requirements of the theory”. On the other hand *in fact* would mean: “with respect to experiment”.

It is very difficult to describe formally the notion of *complementarity*. In order to be complementary, two observables must be non-simultaneously measurable by right. That condition can be taken as a formal description of complementarity; hence a theory with complementarity is a theory including non-simultaneously measurable observables.

6°) A theory is *deterministic* by right if there exists at least one initial

element X_0 such that from this element X_0 it is possible to predict with certainty the value of all observables at any time. A theory is called *essentially indeterministic* if it does not contain such an X_0 . It can be proved that a subjectivistic theory is essentially indeterministic, and that an essentially indeterministic theory (i.e. a theory with indeterminism by right) is a subjectivistic theory [10b, pp. 241–244, 260–284].

7°) In a subjectivistic theory, it is necessary to use an apparatus in order to obtain some information on the observed physical system, and that apparatus cannot be eliminated from the theoretical description. Conversely if the use of an apparatus is essential by right (and not only in fact) the theory is subjectivistic.

The preceding notions can be defined more precisely; to each physical notion corresponds a definite term in the formal description of the physical facts, that is in the formalism of the prediction calculus; such definitions bring in, in a precise way, the properties pointed out here.

From these definitions it follows that the uniqueness of the f -function and the form imposed by the spectral decomposition theorem is a consequence of only one of the following assumptions, and any one of them implies the others:

- 1) the theory is a subjectivistic one,
- 2) there is no state-observable,
- 3) an experimental datum is not an intrinsic property of the observed physical system,
- 4) imprecise experimental data cannot be analysed,
- 5) there are two observables not simultaneously measurable,
- 6) there is some complementarity,
- 7) there is essential indeterminism,
- 8) by right it is necessary to use an apparatus in order to obtain some information on the observed physical system.

This last condition is the most intuitive for microphysics and it can be placed as postulate under the form of *principle of observability* [13; 10b, pp. 316–318].

On the contrary, if we assume the negation of one of the above assumptions, this implies the negation of the others and the prediction scheme reduces to a phase-space scheme. These conditions have as a consequence that the observable physical systems can be divided into two classes:

- 1) systems which are, by right, directly observable by means of the

sense organs of the observers (i.e. systems for which all observables are simultaneously measurable by right).

2) systems which, by right, can only be observed indirectly by means of certain systems of the preceding class called "apparatus" (i.e. systems in which there exists at least one pair of non-simultaneously measurable observables).

14. The principle of evolution. In the general formalism of our prediction calculus, the evolution of the observed physical system S is described only by the \mathfrak{U} -evolution operator. Any condition concerning the evolution of S consists in a condition assigned to $\mathfrak{U}(t, t_0)$.

To determine the evolution of this \mathfrak{U} -operator it is natural to admit the following principle as a fundamental property for predictions [14]: "If during the time interval (t_0, t) no measurement is realised on the observed physical system S , (an initial measurement being made at t_0), then the prediction for an instant τ (between t_0 and t) has an effect upon the predictions for the instant t , and this for all τ ".

Any prediction for the instant τ is obtained from a predictional-element $X(\tau)$ and $X(\tau) = \mathfrak{U}(\tau, t_0)X_0$. A prediction for the instant t is calculated from the predictions for different times between t_0 and t . That is, any prediction for an instant τ is considered as an indication for computing a prediction for the instant t . A prediction for the instant τ is computed from $X(\tau)$ (by the spectral decomposition theorem); in other words this indication is described by $X(\tau)$, and the contribution of $X(\tau)$ in order to calculate $X(t)$ is an element $Y(t, \tau)$ obtained as a function of $X(\tau)$, that is

$$Y(t, \tau) = \mathfrak{F}_*(t, \tau)X(\tau)$$

where $\mathfrak{F}_*(t, \tau)$ is an operator.

Considering $n + 1$ instants

$$\tau_0 = t_0, \tau_1, \tau_2, \dots, \tau_i, \dots, \tau_{n-1}, \tau_n = t$$

we shall have

$$X(t) = \sum_{i=0}^n \mathfrak{F}_n(t, \tau_i)X(\tau_i)\Delta\tau_i$$

The process used to define an integral gives us

$$X(t) = X_0(t) + \int_{t_0}^t \mathfrak{F}(t, \tau)X(\tau)d\tau$$

where

$$X_0(t) = \lim_{n \rightarrow \infty} \mathfrak{F}_n(t, t_0) X_0 \cdot \Delta \tau_0$$

This is a functional equation for $X(t)$, we have

$$X(t) = \mathfrak{U}(t, t_0) X_0,$$

hence

$$\mathfrak{U}(t, t_0) = \mathfrak{V}(t, t_0) + \int_{t_0}^t \mathfrak{F}(t, \tau) \mathfrak{U}(\tau, t_0) d\tau$$

with $\mathfrak{V}(t, t_0) X_0 = X_0(t)$.

The equation for the operator $\mathfrak{U}(t, t_0)$ has the form of a Volterra's integral equation of hereditary process, but it is an equation between operators and not an equation between functions.

If $\mathfrak{V}(t, t_0)$ has the properties of an evolution operator, it can be interpreted as the evolution operator of a fictive system S_0 called a *substratum* for S . Also S can be interpreted as a perturbed system and S_0 as a non perturbed system. The equation in \mathfrak{U} can be solved by a process of successive approximations; the first step gives the usual perturbation of first order and the upper steps the perturbations of higher orders [18].

In the general case, \mathfrak{U} is not derivable and there is no Hamiltonian, and thus no wave equation; but in many particular cases, \mathfrak{U} has a time derivative and obeys a differential equation:

$$\frac{\partial \mathfrak{U}}{\partial t} = -i\hbar \mathfrak{S} \mathfrak{U}$$

where \mathfrak{S} is an operator called the *Hamiltonian*.

We have

$$\mathfrak{S} = \mathfrak{S}_0 + \frac{i}{\hbar} \mathfrak{F}(t, t)$$

if $\mathfrak{V}(t, t_0)$ obeys an equation of this form. We have the wave equation if \mathfrak{U}_0 and $\mathfrak{F}(t, \tau)$ have a time derivative and if $\mathfrak{F}(t, \tau)$ tends to a limit when τ tends to t . But in general $\mathfrak{F}(t, \tau)$ does not tend to a limit when τ tends to t and there is only the integral operatorial equation to describe the evolution of the system.

15. Experimental sentences. The general theory of predictions leads us to single out sentences of a special type: the experimental sentences on

which a calculus can be defined. Thus we get an algebra, which plays an important part in the physical theories under consideration [15; 10b, pp. 91–215].

16. Search for new theories. Physico-logical studies alone do not allow us to build a new physical theory [16; 4c, pp. 54–60]. A new theory can only be obtained by thoroughly deepening the meaning of the purely physical notions of a theory. But physico-logical studies definitely help us. For example, in the recent discussions about the quantum theories, concerning the discrepancy between the statistical interpretation and the causal one, the physico-logical considerations served to yield precise answers: if we have an essentially indeterministic theory, it is always possible to construct a deterministic theory which gives us the same results (i.e. the same predictions concerning future measurements) under the following conditions: i) the notion of a physical system is not the same in both theories, ii) we must add hidden parameters, some belonging to the physical system (in the sense of the indeterministic theory) and some to the measuring apparatus; moreover some of these hidden parameters are not measurable in any way (they are metaphysical parameters) [8; 12c, pp. 43–100]. Reciprocally P. Février has proved [17; 12c, pp. 135–150] that, if we have a deterministic theory with hidden parameters, by eliminating these parameters and modifying the notion of a physical system, we obtain an essentially indeterministic theory. Hence the notions of determinism and indeterminism are not physical notions, properties of nature, but are relative to the theoretical requirements.

17. Various levels. Whereas, in the study of mathematical theories, it is enough to distinguish two levels: the theoretical one, and the metatheoretical one, or in other words, the language and the metalanguage, in the study of physical theories, we have to distinguish a greater number of levels: for instance, the language of the experimental sentences, the language of predictions, the language of the theory, the metalanguage [15].

18. Various approaches. Physico-logical studies are still little developed, and many problems are to be formulated. To end, I shall point out the main approaches as follows:

a) To study in a strictly logical way a given physical theory only taken as a deductive theory.

b) To elaborate general physico-logical considerations when a connexion with experiment is introduced by the notion of adequacy.

c) To come to more particular physico-logical considerations when the formal structure of a theory is taken into account.

d) To draw the consequences of the following notions: measurements, experimental statements, predictions. That is to say: to work out the general theory of predictions.

e) To study the calculus of experimental sentences and enter into epitheoretical considerations about the general theory of predictions.

f) In particular, the physico-logical researches allow us to separate in a physical theory the intrinsic (or objectivitic) properties of the physical objects, from those which are intrinsic properties of the compound object-apparatus, but not of the objects themselves. Criteria for the intrinsic and extrinsic properties have been mentioned.

These considerations on intrinsic and non-intrinsic properties played an important part in the recent developments of physical theories, namely in the elaboration of the functional theory of particles. In this theory, a particle is no longer described by a point, but by a function u or a finite set of functions u_i . I have not space enough here to give details about this theory which I have developed in recent papers [9].

19. Conclusion. The modern physical theories involve such various and mixed levels of thought that, besides purely physical, logical and mathematical considerations, they need intermediate researches in order to connect together these different kinds of developments.

Physico-logic is such an intermediate field, and that is the reason why physico-logical methods do not quite fulfill the formal conditions required either from a physical theory or from a logical one. But one cannot hope to surmount the present heavy difficulties of theoretical physics only by means of the formal achievement of reasonings. Adequacy has to be realised first of all by a physical theory and, for that purpose, physico-logical studies can be very helpful and set the theoretical developments in their right connection with experiment. They are presently in their first stage, like the studies about foundations of mathematics at their beginning; the formal achievement does not appear at the beginning, it depends on the efficiency of the methods under consideration, and, on the other hand, their efficiency depends on their formal strictness. Physico-logical studies must be broadly developed in both directions, and play an important part in the future.

Bibliography

- [1] AESCHLIMANN, F., a) *Sur la persistance des structures géométriques dans le développement des théories physiques*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 232 (1951), pp. 695-597.
 b) *Recherches sur la notion de système physique*. Thèse de Doctorat ès-Sciences, Paris 1957.
- [2] — and J. L. DESTOUCHES, *L'électromagnétisme non linéaire et les photons en théorie fonctionnelle des corpuscules*. Journal de Physique et le Radium, t. 18 (1957), p. 632.
- [3] CAZIN, M., a) *Algorithmes et théories physiques*. Comptes Rendus des séances de l'Académie des Sciences de Paris, t. 224, pp. 541-543.
 b) *Algorithmes et construction d'une théorie unifiante*. Comptes Rendus des séances de l'Académie des Sciences, t. 224 (1947), pp. 805-807.
 c) *Persistance des structures formelles dans le développement des théories physiques*. Thèse de Doctorat Univ. Paris, Lettres-Philosophie, Paris 1947.
 d) *Les structures formelles des mécaniques ondulatoires et leur persistance dans les nouvelles tentatives théorique*. Thèse de Doctorat ès-Sciences, Paris 1949.
- [4] DESTOUCHES, J. L., a) *Essai sur la forme générale des théories physiques*. Thèse principale pour le Doctorat ès-Lettres, Paris 1938. Monographies mathématiques de l'Université de Cluj, fasc. VII, Cluj (Roumanie) 1938.
 b) *Principes fondamentaux de Physique théorique*. Vol. 1, Paris 1942, 174 + IV pp.
 c) *Traité de physique théorique et de physique mathématique*, t. I. *Méthodologie, Notions géométriques*, vol. I, Paris 1953, 228 + XIV pp.
- [5] —, a) *Unité de la physique théoriques*. Comptes Rendus des séances de l'Académie des sciences de Paris, vol. 205 (1947), pp. 843-845.
 b) *Essai sur l'Unité de la physique théorique*. Thèse complémentaire pour le Doctorat ès-Lettres, Paris 1938; Bulletin scientifique de l'Ecole polytechnique de Timisoara, Roumanie 1938.
- [6] — a) *Les espaces abstraits en Logique et la stabilité des propositions*. Bulletin de l'Académie royale de Belgique (classe des sciences) 5^e sér., vol. XXI (1935), pp. 780-86.
 b) *Le rôle de la notion de stabilité en physique*. Bulletin de l'Académie royale de Belgique (classe des sciences) 5^e sér., vol. XXII (1936), pp. 525-532.
 c) *Conditions minima auxquelles doit satisfaire une théorie physique*. Bulletin de l'Académie royale de Belgique (classe des sciences) 5^e sér., vol. XXIII (1937), pp. 159-165.
 d) *Loi générale d'évolution d'un système physique*. Journal de Physique et le Radium, sér. 7, vol. 7 (1936), pp. 305-311.
 e) *La notion de grandeur physique*. Journal de Physique et le Radium, sér. 7, vol. 7 (1936), pp. 354-360.
 f) *Le principe de Relativité et la théorie générale de l'évolution d'un système physique*. Journal de Physique et le Radium, sér. 7, vol. 7 (1936), pp. 427-433.
 g) *Les prévisions en physique théorique*. Communication au Congrès inter-

- national de Philosophie des Sciences, Octobre 1949, Actualités scientifiques et industrielles Hermann, Paris 1949.
- h) *Corpuscules et Systèmes de Corpuscules, Notions fondamentales*. Vol. 1, Paris 1941, 342 pp.
- i) *Principes fondamentaux de physique théorique*. Vol. II, Paris 1942, 484 + VI pp.; vol. III, Paris 1942, 248 + IV pp.
- j) *Über den Aussagenkalkül der Experimentalaussagen*. Archiv für mathematische Logik und Grundlagenforschung, Heft 2/2-4, pp. 424-25.
- [7] —, *Cours mimeogr.* Faculté des Sciences, Paris 1957.
- [8] —, a) *Sur l'interprétation physique de la Mécanique ondulatoire et l'hypothèse des paramètres cachés*. Journal de Physique et le Radium, vol. 13 (1952), pp. 354-358.
- b) *Sur l'interprétation physique des théories quantiques*. Journal de Physique et le Radium, vol. 13 (1952), pp. 385-391.
- [9] —, a) *Funktionelle Theorie der Elementarteilchen*. Vorlesung Pariser Universitätswoche, München 1955, pp. 176-183.
- b) *Fonctions indicatrices de spectres*. Journal de Physique et le Radium, vol. 17 (1956), p. 475.
- c) *Quantization in the functional theory of particles*. Nuovo Cimento, suppl. vol. III, sér X (1956), pp. 433-468.
- d) *La quantification en théorie fonctionnelle des corpuscules*. Vol. 1, Paris 1956, VI + 144 pp.
- e) *Le graviton et la gravitation en théorie fonctionnelle des corpuscules*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 245 (1957), pp. 1518-1520.
- f) *La gravitation en théorie microphysique non linéaire*. Journal de Physique et le Radium, vol. 18 (1957), p. 642.
- g) *Le graviton en théorie fonctionnelle des corpuscules*. Journal de Physique et le Radium, vol. 19 (1958), pp. 135-139.
- h) Journal de Physique et le Radium, vol. 19 (1958) (sous presse)
- i) *Corpuscules et champs en théorie fonctionnelle*. vol. 1, Paris 1958, VIII + 164 pp.
- j) *Les systèmes de corpuscules en théorie fonctionnelle* (A.S.L. Hermann, Paris 1958).
- [10] FEVRIER, P., a) *Recherches sur la structure des théories physiques*. Thèse Sciences Math. Univers., Paris 1945.
- b) *La structure des théories physiques*. Paris, 1951, XII + 424 pp.
- c) *Logical Structure of Physical Theories*. This volume.
- [11] —, a) *Déterminisme et indéterminisme*. Vol. 1, Paris 1955, 250 pp.
- b) *L'interprétation physique de la Mécanique ondulatoire et des théories quantiques*. Vol. 1, Paris 1956, 216 pp.
- c) *Determinismo e indeterminismo*. Vol. 1, Mexico 1957, 270 pp.
- [12] —, a) *Signification profonde du principe de décomposition spectrale*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 222 (1946), pp. 866-868.
- b) *Sur l'interprétation physique de la Mécanique ondulatoire*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 222 (1946), p. 1087.

- c) *L'interprétation physique de la Mécanique ondulatoire et des théories quantiques*. vol. 1, Paris 1956, 216 pp.
- [13] —, *Monde sensible et monde atomique*. Theoria (Philosophical Miscellany presented to Alf Nyman), 1949, pp. 79–88.
- [14] —, a) *Sur la recherche de l'équation fonctionnelle d'évolution d'un système en théorie générale des prévisions*. Comptes Rendus des séances de l'Académie des Sciences de Paris, Vol. 230 (1950), pp. 1742–1744.
b) *Sur la notion de système physique*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 233 (1959), p. 604.
- [15] —, *La logique des propositions expérimentales*. Actes du 2° colloque de Logique mathématique de Paris 1952, Paris 1954, pp. 115–118.
- [16] —, a) *Sur la notion d'adéquation et le calcul minimal de Johansson*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 224 (1947), pp. 545–548.
b) *Adequation et développement dialectique des théories physiques*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 224 (1947), pp. 807–810.
- [17] —, *Sur l'élimination des paramètres cachés dans une théorie physique*. Journal de Physique et le Radium, vol. 14 (1953), p. 640.
- [18] GUY, R., a) Comptes Rendus séances de l'Académie des Sciences de Paris, 1950–1953.
b) Thèse de Doctorat ès-Sciences mathématiques, Univ. Paris 1954.
- [19] NIKODYM, O. M., *Remarques sur les intégrales de M. J. L. Destouches considérées dans sa théorie des prévisions*. Comptes Rendus des séances de l'Académie des Sciences de Paris, vol. 225 (1947), p. 479.

PART III
GENERAL PROBLEMS AND APPLICATIONS
OF THE AXIOMATIC METHOD

STUDIES IN THE FOUNDATIONS OF GENETICS

J. H. WOODGER

University of London, London, England

In what follows a fragment of an axiom system is offered — a fragment because it is still under construction. One of the ends in view in constructing this system has been the disclosure, as far as possible, of *what is being taken for granted* in current genetical theory, in other words the discovery of the hidden assumptions of this branch of biology. In the following pages no attempt will be made to give a comprehensive account of all the assumptions of this kind which have so far been unearthed; attention will be chiefly concentrated on one point — the precise formulation of what is commonly called Mendel's First Law, and its formal derivation from more general doctrines, no step being admitted only because it is commonly regarded as intuitively obvious. Mendel's First Law is usually disposed of in a few short sentences in text-books of genetics, and yet when one attempts to formulate it quite explicitly and precisely a considerable wealth and complexity of hidden assumptions is revealed. Another and related topic which can be dealt with by the axiomatic method is the following. Modern genetics owes its origin to the genius of Mendel, who first introduced the basic ideas and experimental procedures which have been so successful. But it is time to inquire how far the Mendelian hypotheses may now be having an inhibiting effect by restricting research to those lines which conform to the basic assumptions of Mendel. It may be profitable to inquire into those assumptions in order to consider what may happen if we search for regions in which they do not hold. The view is here taken that the primary aim of natural science is discovery. Theories are important only in so far as they promote discovery by suggesting new lines of research, or in so far as they impose an order upon discoveries already made. But what constitutes a discovery? This is not an easy question to answer. It would be easier if we could identify observation and discovery. But the history of natural science shows abundantly that such an identification is impossible. Christopher Columbus sailed west from Europe and returned with a report that he had found land. What made this a discovery was the fact

that subsequent travellers after sailing west from Europe also returned with reports which agreed with that of Columbus. If the entire American continent had quietly sunk beneath the wave as soon as Columbus's back was turned we should not now say that he had discovered America, even although he had observed it. If an astronomer reported observing a new comet during a certain night, but nobody else did, and neither he nor anybody else reported it on subsequent nights, we should not say that he had made a discovery, we should say that he had made a mistake. Observations have also been recorded which have passed muster for a time but have finally been rejected, so that these were not discoveries. Moreover, there have been observations (at least in the biological sciences) which have been ignored for nearly fifty years before they have been recognized as discoveries. Theories play an important part in deciding what is a discovery. Under the influence of the doctrine of preformation, in the early days of embryology, microscopists actually reported seeing little men coiled up inside spermatozoa. Under the influence of von Baer's germ-layer theory the observations of Julia Platt on ecto-mesoderm in the 1890s were not acknowledged as discoveries until well into the twentieth century. Such considerations raise the question: is Mendelism now having a restricting effect on genetical research?

The distinction between records of observations and formulations of discoveries is particularly sharp in genetics; as we see when we attempt to formulate carefully Mendel's observations on the one hand and the discoveries attributed to him on the other. It will perhaps make matters clearer if we first of all distinguish between accessible and inaccessible sets. Accessible sets are those whose members can be handled and counted in the way in which Mendel handled and counted his tall and dwarf garden peas. Inaccessible sets, on the other hand, are those to which reference is usually being made when we use the word 'all'. The set of *all* tall garden peas is inaccessible because some of its members are in the remote past, some are in the (to us) inaccessible future, and some are in inaccessible places. No man can know its cardinal number. But observation records are statements concerning accessible sets and formulations of discoveries are statements concerning inaccessible sets. The latter are therefore hypothetical in a sense and for a reason which does not apply to the former statements. But there are other kinds of statements about inaccessible sets in addition to 'all'-statements. In fact, from the point of view of discoveries, the latter can be regarded as a special case of a more general kind of statement, namely those statements which give expression

to hypotheses concerning the *proportion* of the members of one set, say X , which belong to a second set Y . When that proportion reaches unity we have the special case where *all* X s are Y s. In the system which is given in the following pages the notation ' pY ' is used to denote the set of all classes X which have a proportion p of their members belonging to Y , p being a fraction such that $0 \leq p \leq 1$. This notation can be used in connexion with both accessible and inaccessible sets. In the latter case it is being used to formulate statements which cannot, from the nature of the case, be known to be true. Such a statement may represent a leap in the dark from an observed proportion in an accessible set, or it may be reached deductively on theoretical grounds. In either case the continued use of a particular hypothesis of this kind depends on whether renewed observations continue to conform to it or not. Statistical theory provides us with tests of significance which enable us to decide which of two hypotheses concerning an inaccessible set accords better with a given set of observations made on accessible sub-sets of the said inaccessible set. In the present article we are not concerned with the questions of testing but with those parts of genetical theory which are antecedent to directly testable statements. At the same time it must be admitted that more is assumed in the hypothesis than that a certain inaccessible set contains a proportion of members of another set. As observations take place in particular places, at particular times, must there not be an implicit reference to times and places in the hypotheses concerning inaccessible sets, if such hypotheses are to be amenable to testing against observations? Consider, for example, the hypothesis that half the human children at the time of birth are boys. This would be the case if all children born in one year were boys and all in the next year were girls, and so on with alternate years, provided the same number of children were born in each year. But clearly a more even spread over shorter intervals of time is intended by the hypothesis. Again, there cannot be an unlimited time reference, because according to the doctrine of evolution there will have been a time when no children were born, and if the earth is rendered uninhabitable by radio-activity a time will come when no more children are born. Thus a set which has accessible sub-sets during one epoch may be wholly inaccessible in another.

In what follows no attempt will be made to solve all these difficult problems; we shall follow the usual custom in natural science and ignore them. Attention will be confined to the *one* problem of formulating Mendel's First Law. In the English translation of Mendel's paper of 1866,

which is given in W. BATESON's Mendel's *Principles of Heredity*, Cambridge 1909, we read (p. 338):

Since the various constant forms are produced in *one* plant, or even in *one* flower of a plant, the conclusion appears to be logical that in the ovaries of the hybrids there are as many sorts of egg cells, and in the anthers as many sorts of pollen cells, as there are possible constant combinations of forms, and that these egg and pollen cells agree in their internal composition with those of the separate forms.

In point of fact it is possible to demonstrate theoretically that this hypothesis would fully suffice to account for the development of the hybrids in the separate generations, if we might at the same time assume that the various kinds of egg and pollen cells were formed in the hybrids on the average in equal numbers.

Bateson adds, in a foot-note to the last paragraph: 'This and the preceding paragraph contain the essence of the Mendelian principles of heredity.' It will be shown below that much more must be assumed than is explicitly stated here. L. Hogben, in *Science for the Citizen*, London, 1942, in speaking of Mendel's Second Law mentions the first in the following passage (p. 982):

It is not, however, a law in the same sense as Mendel's First Law, of *segregation*, which we have deduced above, for it is only applicable in certain cases, and as we shall see later, the exceptions are of more interest than the rule.

But surely, Mendel's First Law is also only applicable in certain cases, and if this is not generally recognized it is because the law is never so formulated as to make clear what those cases are. We cannot simply say that if we interbreed *any* hybrids the offspring will follow the same rules as were reported in Mendel's experiments with garden peas, because it would be possible to quote counter-examples. It is hoped that the following analysis will throw some light on this question and that in this case also the exceptions may prove to be of at least as much theoretical interest as the rule. It will be shown that the condition referred to in the second of the above two paragraphs from Mendel's 1866 paper is neither necessary nor sufficient to enable us to derive the relative frequencies of the kinds of offspring obtainable from the mating of hybrids. It is *not sufficient* because it is also necessary to assume (among other things) that the union of the gametes takes place as random. It is *not necessary* because if the random union of the gametes is assumed the required frequencies can be derived without the assumption of equal proportions of the kinds of

gametes. At the same time it will be seen that a number of other assumptions are necessary which are not usually mentioned and thus that a good deal is being taken for granted which may not always be justified.

When we are axiomatizing we are primarily interested in ordering the statements of a theory by means of the relation of *logical consequence*; but where theories of natural science are concerned we are also interested in another relation between statements, a relation which I will call the relation of *epistemic priority*. A theory in natural science is like an iceberg — most of it is out of sight, and the relation of epistemic priority holds between a statement *A* and a statement *B* when *A* speaks about those parts of the iceberg which are out of water and *B* about those parts which are out of sight; or *A* speaks about parts which are only a little below the surface and *B* about parts which are deeper. In other words: *A* is less theoretical, less hypothetical, assumes less than *B*. If *A* is the statement

Macbeth is getting a view of a dagger

and *B* is the statement

Macbeth is seeing a dagger

then *A* is epistemically prior to *B*. Macbeth was in no doubt about *A*, but he was in serious doubt about *B* and his doubts were confirmed when he tried to touch the dagger but failed to get a feel of it. Again, if *A* is the statement

Houses have windows so that people inside can see things

and *B* is the statement

Houses have windows in order to let the light in

then *A* is epistemically prior to *B*.

We not only say that Columbus discovered America, but also that J. J. Thomson discovered electrons. In doing so we are clearly using the word 'discovered' in two distinct senses. What J. J. Thomson discovered in the first sense was what we may expect to observe when an electrical discharge is passed through a rarified gas. He then *introduced* the word 'electron' into the language of physics in order to formulate a hypothesis from which would follow the generalizations of his discoveries concerning rarified gases. It will help to distinguish the two kinds of discoveries if we call statements which are generalizations from accessible sets to inaccessible sets *inductive* hypotheses, and statements which are introduced in

order to have such hypotheses among their logical consequences *explanatory* hypotheses. Then we can say that to every explanatory hypothesis S_1 there is at least one inductive hypothesis S_2 such that S_2 is a consequence of S_1 (or of S_1 in conjunction with other hypotheses) and is epistemically prior to it. Were this not so S_1 would not be testable. But, as we shall see later, it is also possible to have an explanatory hypothesis S_3 and an inductive hypothesis S_2 , which is *not* a consequence of S_3 although it is epistemically prior to it, *both* of which are consequences of the same explanatory hypothesis S_1 . If what you want to say can be expressed just as well by a statement A as by a statement B then, if A is epistemically prior to B , it will (if no other considerations are involved) be better to use A . In what follows I shall try to formulate all the statements concerned in the highest available epistemic priority. Statements concerning parents and offspring only are epistemically prior to statements which also speak about gametes and zygotes; and statements about gametes and zygotes are epistemically prior to statements which speak also about the parts of gametes and zygotes. The further we go from the epistemically prior inductive hypotheses the more we are taking for granted and the greater the possibility of error. The following discussion of Mendel's First Law will be in terms of parents, offspring, gametes, zygotes and environments.

The foregoing remarks may now be illustrated by a brief reference to Mendel's actual experiments. Suppose X and Y are accessible sets of parents. Let us denote the set of all the offspring of these parents which develop in environments belonging to the set E by

$$f_E(X, Y)$$

If all members of X resemble one another in some respect (other than merely all being members of X) and all members of Y resemble one another in some *other* respect (also other than merely all being members of Y), so that the respect in which members of X resemble one another is distinct from that in which members of Y resemble one another, then $f_E(X, Y)$ constitutes an accessible set of *hybrids*. We also need $f_E^2(X, Y)$ which is defined as follows:

$$f_E^2(X, Y) = f_E(f_E(X, Y), f_E(X, Y))$$

Mendel experimented with seven pairs of mutually exclusive accessible sets and the hybrids obtained by crossing them. It will suffice if we consider one pair. Let 'A' denote the pea plants with which Mendel began

his experiments and which were tall in the sense of being about six feet high; and let 'C' denote the peas which he used and which were dwarf in the sense of being only about one foot high. Let us use 'T' to denote the inaccessible set of *all* tall pea plants and 'D' to denote the inaccessible set of *all* dwarf pea plants. Thus we have

$$A \subset T \text{ and } C \subset D$$

let us use 'B' to denote the set of all environments in which Mendel's peas developed. Mendel first tested his As and Cs to discover whether they bred true and found that they did because

$$f_B(A, A) \subset T \text{ and } f_B^2(A, A) \subset T$$

$$f_B(C, C) \subset D \text{ and } f_B^2(C, C) \subset D$$

He next produced hybrids and reported that

$$f_B(A, C) \subset T$$

$$f_B^2(A, C) \in \frac{787}{1064} T \cap \frac{277}{1064} D$$

Finally, he took 100 of the tall members of $f_B^2(A, C)$ and self fertilized them. From 28 he obtained only tall plants and from 72 he obtained some tall and some dwarf. This indicated that about one third of the tall plants of $f_B^2(A, C)$ were pure breeding tall like $f_B(A, A)$ and two thirds were like the hybrid tall or $f_B(A, C)$.

Closely similar results were obtained in the other six experiments, although the respects in which the plants differed were in those cases not concerned with height but with colour or form of seed or pod or the position of the flowers on the stem. In each case the hybrids all resembled only one of the parental types, which Mendel accordingly called the dominant one. The parental type which was not represented in the first hybrid generation, but which reappeared in the second, he called the recessive one. Mendel took the average of the seven experiments and sums up as follows:

If now the results of the whole of the experiments be brought together, there is found, as between the number of forms with the dominant and recessive characters, an average ratio of 2.98 to 1, or 3 to 1.

So long as we assert that the average ratio is 2.98 to 1 we are dealing with accessible sets and have no law or explanatory hypothesis. But what does

Mendel's addition 'or 3 to 1' mean? Presumably these few words express the leap from an observed proportion in an accessible set to a hypothetical proportion in an inaccessible set. This represents Mendel's discovery as opposed to his observations. At the same time there is no proposal to extend this *beyond* garden peas. This extension was done by Mendel's successors who, on the basis of many observations, extended his generalization regarding the proportions of kinds of offspring of hybrids over a wide range of inaccessible sets not only of plants but also of animals. In addition to this Mendel also left us his explanatory hypothesis, the hypothesis namely that the hybrids produce gametes of two kinds — one resembling the gametes produced by the pure dominant parents, and the other resembling those produced by the recessive parents. He also assumed that these two kinds of gametes were produced in equal numbers. We have now to consider what is the minimum theoretical basis for deriving this hypothesis as a theorem in an axiom system.

A GENETICAL AXIOM SYSTEM

(In what follows the axiom system is given in the symbolic notation of set-theory, sentential calculus and the necessary biological functors (the last in bold-face type). Accompanying this is a running commentary in words intended to assist the reading of the system; but it must be understood that this commentary forms no part of the system itself.)

The following primitives suffice for the construction of a genetical axiom system expressed on the level of epistemic priority here adopted; for cyto-genetics (and even perhaps for extending the present system) additional primitives are necessary.

- (i) ' uFx ' for ' u is a gamete which fuses with another gamete to form the zygote (fertilized egg) x '.
- (ii) ' $dlz\ xyz$ ' for ' x is a zygote which develops in the environment y into the life z '.
- (iii) ' $u\ gam\ z$ ' for ' u is a gamete produced by the life z '.
- (iv) δ is the class of all male gametes.
- (v) φ is the class of all female gametes.
- (vi) '**phen**' is an abbreviation for 'phenotype'

The following postulates are needed for the derivation of the theorems which are to follow:

POSTULATE 1 $(u)(v)(x):uFx.vFx.u \neq v.\supset.\sim(\exists w).wFx.w \neq u.w \neq v$

This asserts that not more than two gametes fuse to form each zygote.

POSTULATE 2 $(x)(w):(Eu).uFx.uFw.\supset.x = w$

This asserts that if a gamete unites with another to form a zygote then there is no other zygote for which this is true.

POSTULATE 3 $(u)(v)(x):.uFx.vFx.u \neq v.\supset:u \in \mathfrak{J}.v \in \mathfrak{F}.\vee.u \in \mathfrak{F}.v \in \mathfrak{J}$

This asserts that of the two gametes which unite to form any zygote one is a male gamete and the other a female gamete.

POSTULATE 4 $\mathfrak{J} \cap \mathfrak{F} = \Lambda$

This asserts that no gamete is both male and female.

POSTULATE 5 $(x)(y)(z)(u)(v):dlz\ xyz.dlz\ uvz.\supset.x = y.u = v$

This asserts that every life develops in one and only one environment from one and only one zygote.

POSTULATE 6 $(x)(y)(z)(x')(y')(z'):dlz\ xyz.dlz\ x'y'z'.$

$(\exists u).u\ gam\ z.u\ gam\ z'.\supset.x = x'$

This asserts that if there is a gamete produced by a life z and the same gamete is produced by a life z' then the zygote from which z develops is identical with the zygote from which z' develops. This may seem strange until it is explained that by 'a life' is here meant something with a beginning and an end in time and a fixed time extent. The expression is thus being used in a way somewhat similar to the way in which it is used in connexion with life insurance. Suppose a zygote is formed at midnight on a certain day; suppose it develops for say ten days and on that day death occurs. Then the whole time-extended object of ten days duration 'from fertilization to funeral' is a life which is *complete in time*. But suppose we are only concerned with what happens during the first ten *hours*; then that also is a life, in the sense in which the word is here used, and one which is a proper part of the former one. Now if a gamete is said to be produced by the shorter life it is also produced by the longer one of which that shorter one is a part; we cannot identify the two lives but we can say that they both develop from the same zygote. As here understood the time-length of a life fixes its environment; because the environment of a life is the sphere and its contents which has the zygote from which development begins as its centre and a radius which is equal in light-years to

the length of the life in years. But no time-metric is needed for the present system and many complications are therefore avoided.

All the primitive notions of this system are either relations between individuals or are classes of individuals. But the statements of genetics with which we are concerned in what follows do not speak of individual lives, individual environments, individual zygotes or individual gametes but of *classes* of individuals and of relations between such classes. But the classes we require are definable by means of the primitives.

DEFINITION 1 $x \in U(\alpha, \beta) . \equiv : (\exists u)(\exists v) . u \in \alpha . v \in \beta . u \neq v . uFx . vFx$

We thus use ' $U(\alpha, \beta)$ ' to denote the class of all zygotes which are formed by the union of a gamete belonging to the class α with one belonging to the class β .

DEFINITION 2 $z \in L_E(Z) . \equiv : (\exists x)(\exists y) . dlz\ xyz . x \in Z . y \in E$

' $L_E(Z)$ ' is used to denote the class of all lives which develop from a zygote belonging to the class Z in an environment belonging to the class E .

DEFINITION 3 $u \in G_E(X) . \equiv : (\exists x)(\exists y)(\exists z) . dlz\ xyz . y \in E . z \in X . u\ gam\ z$
' $G_E(X)$ ' thus denotes the class of all gametes which are produced by lives belonging to the class X when they develop in environments belonging to the class E .

DEFINITION 4 $z \in Fil_{K,M,E}(X, Y) . \equiv : (\exists x)(\exists y)(\exists u)(\exists v) . u \in G_K(X) .$

$v \in G_M(Y) . uFx . vFx . u \neq v . y \in E . dlz\ xyz$

The letters '*Fil*' are taken from the word 'filial'. The above definition provides a notation for the class of all offspring which develop in environments belonging to the class E and having one parent belonging to the class X and developing in an environment belonging to the class K and the other parent belonging to the class Y and developing in an environment belonging to the class M . For Mendelian contexts only one environmental class need be considered; provision for this simplification is made below.

The above four definitions suffice for most purposes. But it frequently happens that we need to substitute one of the above expressions for the variables of another and in that way very complicated expressions may arise. In order to avoid this the following abbreviations are introduced by

definition:

DEFINITION 5 $D(\alpha, \beta, E) = L_E(U(\alpha, \beta))$

DEFINITION 6 $G(\alpha, \beta, E) = G_E(L_E(U(\alpha, \beta)))$

DEFINITION 7 $F'_{K,M,E}(\alpha, \beta; \gamma, \delta) = Fil_{K,M,E}(D(\alpha, \beta, K), D(\gamma, \delta, M))$

DEFINITION 8 $F_E(\alpha, \beta; \gamma, \delta) = F'_{E,E,E}(\alpha, \beta; \gamma, \delta)$

All the foregoing notions are general and familiar ones. We must now turn to some of a more special and novel kind. If our present inquiry were not confined to the single topic of Mendel's First Law we should at this stage introduce the notion of a genetical system, and we should maintain that genetical systems as then intended constitute the proper objects of genetical investigations. But for the present purpose it suffices if we speak of a specially simple kind of genetical system which we shall call *genetical units*. A genetical unit is a set of three classes: one is a phenotype, another is a class of gametes and the third is a class of environments; — provided certain conditions are satisfied. Suppose $\{P, \alpha, E\}$ is a candidate for the title of genetical unit; then it must be *developmentally closed*, that is to say $D(\alpha, \alpha, E)$ must be a non-empty class and it must be included in the phenotype P ; next it must be *genetically closed*, that is to say $G(\alpha, \alpha, E)$ must be non-empty and must be included in α . Thus neither the process of development nor that of gamete-formation takes us out of the system; it thus 'breeds true'. The official definition is:

DEFINITION 9 $S \in \text{genunit} \equiv :(\exists P)(\exists \alpha)(\exists E). P \in \text{phen}. S = \{P, \alpha, E\}.$

$$D(\alpha, \alpha, E) \neq \Lambda. D(\alpha, \alpha, E) \subseteq P. G(\alpha, \alpha, E) \neq \Lambda.$$

$$G(\alpha, \alpha, E) \subseteq \alpha$$

The genetical systems with which Mendel worked were genetical units, sums of two genetical units and what may be called set-by-set products of such sums. Thus if $\{P, \alpha, E\}$ and $\{Q, \beta, E\}$ are genetical units with the phenotype P dominant to the phenotype Q we shall have $D(\alpha, \beta, E) \neq \Lambda$ and $D(\alpha, \beta, E) \subseteq P$, so that $\{P, Q, \alpha, \beta, E\}$, the sum of the two units, is developmentally closed; if we also have $G(\alpha, \beta, E) \neq \Lambda$ and $G(\alpha, \beta, E) \subseteq \alpha \cup \beta$, then the sum is also genetically closed. As we shall see shortly these assumptions do not suffice to enable us to infer that the sum will behave according to the Mendelian generalizations. If $\{R, \gamma, E\}$ and $\{S, \delta, E\}$ are two more genetical units so that $\{R, S, \gamma, \delta, E\}$ is their sum, then the

set-by-set product of this sum and the former one will be

$$\{P \cap R, Q \cap R, P \cap S, Q \cap S, \alpha \cap \gamma, \beta \cap \gamma, \alpha \cap \delta, \beta \cap \delta, E\}$$

and if it is developmentally and genetically closed this will constitute yet another type of genetical system which was studied by Mendel and with which his Second Law was concerned.

Before we can proceed with the biological part of our system we must now say something about the set-theoretical framework within which it is being formulated and on the basis of which proofs of theorems are carried out. We begin with two important definitions, one of which has already been mentioned. (The definitions and theorems of this part of the system will have Roman numerals assigned to them in order to distinguish them from biological definitions and theorems).

$$\text{DEFINITION I} \quad X \in pY \equiv \frac{N(X \cap Y)}{N(X)} = p \cdot 0 \leq p \leq 1$$

pY is thus the set of all classes which have a proportion p of their members belonging to Y . $N(X)$ is the cardinal number of the class X .

$$\text{DEFINITION II} \quad Z \in [X, Y] \equiv (\exists u)(\exists v) \cdot u \in X \cdot v \in Y \cdot u \neq v \cdot Z = \{u, v\}$$

$[X, Y]$ is the pair-set of the classes X and Y , that is to say it is the set of all pairs (unordered) having one member belonging to the class X and the other to the class Y .

No attempt is made here to present the set-theoretical background axiomatically. We simply list, for reference purposes, the following theorems which can be proved within (finite) set theory and arithmetic.

$$\text{THEOREM I} \quad N(X) = 0 \equiv X = \Lambda$$

$$\text{THEOREM II} \quad X \subseteq Y \supset N(X) \leq N(Y)$$

$$\text{THEOREM III} \quad N(X \cup Y) = N(X \cap \bar{Y}) + N(X \cap Y) + N(\bar{X} \cap Y)$$

$$\text{THEOREM IV} \quad X \cap Y = \Lambda \supset N(X \cup Y) = N(X) + N(Y)$$

$$\text{THEOREM V} \quad N([X, Y]) = N(X \cap \bar{Y}) \cdot N(\bar{X} \cap Y) + N(X \cap Y).$$

$$[N(X \cap \bar{Y}) + N(\bar{X} \cap Y) + N(X \cap Y) - 1]$$

$$\text{THEOREM VI} \quad X \cap Y = \Lambda \supset N([X, Y]) = N(X) \cdot N(Y)$$

$$\text{THEOREM VII} \quad X \neq \Lambda \cdot X \subseteq Y \equiv X \in 1Y$$

THEOREM VIII $X \in pY \cap qY. \supset. p = q$

THEOREM IX $X \neq \Lambda. \supset. (\exists p). \frac{N(X \cap Y)}{N(X)} = p. 0 \leq p \leq 1$

THEOREM X $X \neq \Lambda. X \subseteq Y \cup Z. Y \cap Z = \Lambda. \equiv. (\exists p).$

$$X \in pY \cap (1 - p)Z$$

THEOREM XI $X \cap Y = \Lambda. \supset. (pX \cap qY) \subseteq (p + q)(X \cup Y)$

THEOREM XII $Y \subseteq P. Z \subseteq Q. P \cap Q = \Lambda. \supset. (pY \cap (1 - p)Z)$

$$\subseteq (pP \cap (1 - p)Q)$$

THEOREM XIII $Y \subseteq A. Z \subseteq B. W \subseteq C. A \cap B = B \cap C = C \cap A =$

$$= \Lambda. \supset. (pY \cap qZ \cap (1 - p - q)W) \subseteq (pA \cap qB \cap (1 - p - q)C)$$

THEOREM XIV $A \subseteq P. B \subseteq P. C \subseteq Q. A \cap B = B \cap C = C \cap A = P \cap Q = \Lambda. \supset.$

$$(pA \cap qB \cap (1 - p - q)C) \subseteq (p + q)P \cap (1 - p - q)Q$$

THEOREM XV $X \cap Y = \alpha \cap \beta = \Lambda. X \in p\alpha \cap (1 - p)\beta. Y \in q\alpha \cap (1 - q)\beta.$

$$\supset. [X, Y] \in pq[\alpha, \alpha] \cap (p(1 - q) + q(1 - p))$$

$$[\alpha, \beta] \cap (1 - p)(1 - q)[\beta, \beta]$$

THEOREM XVI $X \cap Y = \alpha \cap \beta = \Lambda. \supset. X \in \frac{1}{2}\alpha \cap \frac{1}{2}\beta. Y \in \frac{1}{2}\alpha \cap \frac{1}{2}\beta. \equiv.$

$$N(X \cap \beta, Y \cap \alpha) = N(X \cap \alpha, Y \cap \beta). X \subseteq \alpha \cup \beta. Y \subseteq \alpha \cup \beta$$

We can now return to the biological part of our system. In genetical statements the notion of randomness frequently occurs. It will be required in two places in the present context. In both of these it means persistence of certain relative frequencies during a process. It means the absence of selection or favouritism.

We shall say that a set S which is the sum of two genetical units is *random with respect to U* or that *the union of the gametes is random in S* if and only if X and Y being any classes of gametes of the form $G(\alpha, \beta, E)$, α and β being any gamete classes and E the environment class of S , whenever we have

$$[X, Y] \in p[\gamma, \delta]$$

we also have

$$U(X, Y) \in pU(\gamma, \delta)$$

γ and δ also being gamete-classes of S . The following definition covers

cases where S has an additional phenotype because there is no dominance.

DEFINITION 10 $S \in \textbf{rand } U. \equiv :(\alpha)(\beta)(\gamma)(\delta)(\zeta)(\theta):(E)(\phi)(\exists S_1)(\exists S_2):$
 $S_1, S_2 \in \textbf{genunit}. S = S_1 \cup S_2. \vee. (\exists R). R \in \textbf{phen}.$
 $S = S_1 \cup S_2 \cup \{R\}. \alpha, \beta, \gamma, \delta, \zeta, \theta, E \in S. [G(\alpha, \beta, E),$
 $G(\gamma, \delta, E)] \in \phi[\zeta, \theta]. \supset.$
 $U(G(\alpha, \beta, E), G(\gamma, \delta, E)) \in \phi U(\zeta, \theta)$

Analogously we can say that such a set S is *random with respect to* $D(E)$ or that *development in members of the environment class* E *of* S *is random* if and only if, whenever we have

$$U(G(\alpha, \beta, E), G(\gamma, \delta, E)) \in \phi U(\zeta, \theta)$$

we also have

$$D(G(\alpha, \beta, E), G(\gamma, \delta, E), E) \in \phi D(\zeta, \theta, E)$$

the Greek letters all being variables whose values are the gamete classes of S_1 and ' E ' being a variable whose single value (in Mendelian cases) is the environmental class of S .

DEFINITION 11 $S \in \textbf{rand } D(E). \equiv :.(\alpha)(\beta)(\gamma)(\delta)(\zeta)(\theta)(E)(\phi):(\exists S_1)(\exists S_2):$
 $S_1, S_2 \in \textbf{genunit}. S = S_1 \cup S_2. \vee. (\exists R). R \in \textbf{phen}.$
 $S = S_1 \cup S_2 \cup \{R\}. \alpha, \beta, \gamma, \delta, \zeta, \theta, E \in S. U(G(\alpha, \beta, E),$
 $G(\gamma, \delta, E)) \in \phi U(\zeta, \theta). \supset. D(G(\alpha, \beta, E), G(\gamma, \delta, E), E) \in$
 $\phi D(\zeta, \theta, E)$

We now give a list of biological theorems which are provable from the postulates and definitions and are used in the proofs of the major theorems to follow. On the right hand side of each theorem are indicated the postulates (P), definitions (D) or theorems (T) required for its proof.

- THEOREM 1** $U(\alpha, \beta) = U(\beta, \alpha)$ [D1.
THEOREM 2 $U(X, X) = U(X \cap \mathfrak{J}, X \cap \mathfrak{F})$ [D1, P3.
THEOREM 3 $\alpha \cap \beta = \Lambda. \supset. U(\alpha, \alpha) \cap U(\beta, \beta) = \Lambda$ [P1, D1.
THEOREM 4 $\alpha \cap \beta = \Lambda. \supset. U(\alpha, \alpha) \cap U(\alpha, \beta) = \Lambda$ [P1, D1.
THEOREM 5 $E \cap K = \Lambda. \vee. Z \cap W = \Lambda. \supset. L_E(Z) \cap L_K(W) = \Lambda$ [D2, P2.

$$\begin{aligned}\text{THEOREM 6 } U(\alpha, \alpha) \cap U(\beta, \beta) &= \\ &= \Lambda. \supset. D(\alpha, \alpha, E) \cap D(\beta, \beta, E) = \Lambda \quad [\text{D5, D2, P2.}]\end{aligned}$$

$$\begin{aligned}\text{THEOREM 7 } U(\alpha, \alpha) \cap U(\alpha, \beta) &= \\ &= \Lambda. \supset. D(\alpha, \alpha, E) \cap D(\alpha, \beta, E) = \Lambda \quad [\text{D5, T5.}]\end{aligned}$$

$$\text{THEOREM 8 } \alpha \cap \beta = \Lambda. \supset. D(\alpha, \alpha, E) \cap D(\beta, \beta, E) = \Lambda \quad [\text{T3, T6.}]$$

$$\text{THEOREM 9 } \alpha \cap \beta = \Lambda. \supset. D(\alpha, \alpha, E) \cap D(\alpha, \beta, E) = \Lambda \quad [\text{T4, T7}]$$

$$\text{THEOREM 10 } D(X \cap \mathfrak{J}, X \cap \mathfrak{Q}, E) = D(X, X, E) \quad [\text{D5, T2.}]$$

$$\begin{aligned}\text{THEOREM 11 } \alpha \cap \beta &= \Lambda. \supset. G(\alpha, \beta, E) \cap G(\beta, \beta, E) = \Lambda \quad [\text{D6, D3, D2,} \\ &\quad \text{P5, P6, T4.}]\end{aligned}$$

$$\text{THEOREM 12 } G(\alpha, \alpha, E) \subseteq \alpha. \supset.$$

$$D(G(\alpha, \alpha, E), G(\alpha, \alpha, E), E) \subseteq D(\alpha, \alpha, E) \quad [\text{D1, D2, D5.}]$$

$$\text{THEOREM 13 } F_E(\alpha, \beta; \gamma, \delta) =$$

$$D(G(\alpha, \beta, E), G(\gamma, \delta, E), E) \quad [\text{D8, D7, D4, D5, D6, D1, D2}]$$

$$\text{THEOREM 14 } \{P, \alpha, E\} \in \text{genunit}. \supset. F_E(\alpha, \alpha; \alpha, \alpha) \subseteq P \quad [\text{T13, D9, T12.}]$$

By a *mating description* is meant a statement of the form $X \subseteq Y$ or $X \in \mathfrak{p}Y$ where 'X' is an expression denoting a set of offspring, e.g. ' $F_E(\alpha, \beta; \alpha, \beta)$ ' and 'Y' denotes a phenotype. We turn now to the task of discovering what must be assumed in order to derive the characteristic Mendelian mating descriptions, beginning with that which asserts the relative frequencies of dominants and recessives in the offspring of hybrids when these are mated with one another. For reference purposes it will be convenient if we use abbreviations for groups of the various separate hypotheses which enter into the antecedents of the following theorems. Let us therefore put:

H 1. for: $\{P, \alpha, E\}, \{Q, \beta, E\} \in \text{genunits}. P \cap Q = \alpha \cap \beta = \Lambda$ ($\{P, \alpha, E\}$ and $\{Q, \beta, E\}$ are genetical units and P and Q , and α and β , are mutually exclusive)

H 2. for: $D(\alpha, \beta, E) \subseteq P$

(the hybrids are included in the phenotype P)

H 3. for: $(\exists R). R \in \text{phen}. D(\alpha, \beta, E) \subseteq R$

(this covers cases where there is no dominance but the hybrids are

included in a third phenotype R)

H 4a. for: $G(\alpha, \beta, E) \cap \delta \in \frac{1}{2}\alpha \cap \frac{1}{2}\beta. G(\alpha, \beta, E) \cap \varphi \in \frac{1}{2}\alpha \cap \frac{1}{2}\beta$

(This is one form of Mendel's own hypothesis. He assumed that in the gametes of the hybrids the two kinds occurred in equal numbers both in the case of male and in the case of female gametes. Theorem XVI shows that the above form is equivalent to this).

H 4b. for: $G(\alpha, \beta, E) \cap \delta \neq \Lambda. G(\alpha, \beta, E) \cap \delta \subseteq \alpha \cup \beta.$

$G(\alpha, \beta, E) \cap \varphi \neq \Lambda. G(\alpha, \beta, E) \cap \varphi \subseteq \alpha \cup \beta$

(This is a weaker form of H 4a because it only assumes non-emptiness and inclusion).

H 4c. for: $G(\alpha, \beta, E) \neq \Lambda. G(\alpha, \beta, E) \subseteq \alpha \cup \beta$

(This is weaker still because it does not make separate statements regarding the gametes of different sex).

H 5. for: $S = \{P, \alpha, E\} \cup \{Q, \beta, E\}. S \in \text{rand } D \cap \text{rand } U(E)$

(This is the hypothesis that the system in question is the sum of two genetical units (H 1) and is random both with respect to the union of the gametes and also with respect to the development of the resulting zygotes in the environments belonging to E).

H 5a. for: $S = \{P, \alpha, E\} \cup \{Q, \beta, E\} \cup \{R\}$ and $S \in \text{rand } U \cap \text{rand } D(E)$

(This is to cover the cases when there is no dominance).

The following theorems are asserted for all values of the variables $P, Q, R, \alpha, \beta, E$.

THEOREM 15 states that if we have H 1, H 2, H 4a and H 5 we also have three quarters of the offspring of the hybrids belonging to the dominant and the remaining quarter to the recessive phenotype.

THEOREM 15 H 1. H 2. H 4a. H 5. $\supset. F_E(\alpha, \beta; \alpha, \beta) \in \frac{3}{4}P \cap \frac{1}{4}Q$

In order to make all the steps explicit we give the following derivation of this theorem:

(1) Using ' X ' as an abbreviation of ' $G(\alpha, \beta, E)$ ' we have, by H 1 and P 4

$$(X \cap \delta) \cap (X \cap \varphi) = \alpha \cap \beta = \Lambda$$

(2) By (1), H 4a and T XV we can write:

$$[X \cap \delta, X \cap \varphi] \in \frac{1}{2} \cdot \frac{1}{2}[\alpha, \alpha] \cap 2(\frac{1}{2} \cdot \frac{1}{2})[\alpha, \beta] \cap \frac{1}{2} \cdot \frac{1}{2}[\beta, \beta]$$

(3) From (2), H 5 and D 10 we are now able to obtain:

$$U(X \cap \mathfrak{J}, X \cap \mathfrak{F}) \in \frac{1}{4}U(\alpha, \alpha) \cap \frac{1}{2}U(\alpha, \beta) \cap \frac{1}{4}U(\beta, \beta)$$

(4) We next obtain from (3), H 5 and D 11:

$$D(X \cap \mathfrak{J}, X \cap \mathfrak{F}, E) \in \frac{1}{4}D(\alpha, \alpha, E) \cap \frac{1}{2}D(\alpha, \beta, E) \cap \frac{1}{4}D(\beta, \beta, E)$$

(5) From H 1, D 9 and H 2 we have:

$$D(\alpha, \alpha, E) \subseteq P \text{ and } D(\alpha, \beta, E) \subseteq P \text{ and } D(\beta, \beta, E) \subseteq Q$$

(6) From H 1 we have $\alpha \cap \beta = \Lambda$ and so with the help of T 8 and T 9 we get:

$$\begin{aligned} D(\alpha, \alpha, E) \cap D(\alpha, \beta, E) &= D(\alpha, \beta, E) \cap D(\beta, \beta, E) = \\ &= D(\beta, \beta, E) \cap D(\alpha, \alpha, E) = \Lambda \end{aligned}$$

(7) From (5) and (6) with the help of T XIV we now get:

$$\frac{1}{4}D(\alpha, \alpha, E) \cap \frac{1}{2}(D(\alpha, \beta, E) \cap \frac{1}{4}D(\beta, \beta, E)) \subseteq \frac{3}{4}P \cap \frac{1}{4}Q$$

(8) By T 10 we have:

$$D(X \cap \mathfrak{J}, X \cap \mathfrak{F}, E) = D(X, X, E)$$

(9) From (4), (7) and (8) we obtain:

$$D(X, X, E) \in \frac{3}{4}P \cap \frac{1}{4}Q$$

(10) Putting ' $G(\alpha, \beta, E)$ ' for ' X ' in (9) in accordance with (1):

$$D(G(\alpha, \beta, E), G(\alpha, \beta, E), E) \in \frac{3}{4}P \cap \frac{1}{4}Q$$

(11) By substitution of ' α ' for ' γ ' and ' β ' for ' δ ' in T 13 we get:

$$F_E(\alpha, \beta; \alpha, \beta) = D(G(\alpha, \beta, E), G(\alpha, \beta, E), E)$$

(12) Finally from (10) and (11) we obtain the required result:

$$F_E(\alpha, \beta; \alpha, \beta) \in \frac{3}{4}P \cap \frac{1}{4}Q$$

Before commenting on this we shall give the remaining theorems.

THEOREM 16 is concerned with the offspring of hybrids when mated with the recessive parents; a mating type commonly called a back-cross. It is stated here in a somewhat unusual form and with the weakest possible antecedent. It states that if the hypotheses H 1, H 2, H 4c and H 5 are adopted then we should expect the proportions of the two pheno-

types in the offspring to be identical with the proportions of the two kinds of gametes in the gametes produced by the hybrids. If, therefore, we assume, on the basis of samples, that $F_E(\beta, \beta; \alpha, \beta) \in \frac{1}{2}P \cap \frac{1}{2}Q$ we must also assume that $G(\alpha, \beta, E) \in \frac{1}{2}\alpha \cap \frac{1}{2}\beta$. The first of these hypotheses is epistemically prior to the second and yet they both occur together in the consequent of this theorem.

THEOREM 16 $H\ 1.H\ 2.H\ 4c.H\ 5.\supset.(\exists p).F_E(\beta, \beta; \alpha, \beta) \in pP \cap (1-p)Q.$
 $G(\alpha, \beta, E) \in p\alpha \cap (1-p)\beta$

The derivation of Theorem 16 requires: T 13, T 11, T X, T XV, T VII, D 9 D 10, D 11 and T XII.

In the next theorem we have the same antecedent as in Theorem 15 except that nothing is assumed about the relative proportions of the two kinds of gamete in the gametes produced by the hybrids.

THEOREM 17 $H\ 1.H\ 2.H\ 4b.H\ 5.\supset.(\exists p)(\exists q).F_E(\alpha, \beta; \alpha, \beta) \in$
 $(p - pq + q)P \cap (1 - p)(1 - q)Q.$
 $G(\alpha, \beta, E) \cap \mathcal{J} \in p\alpha \cap (1 - p)\beta. G(\alpha, \beta, E) \cap \mathcal{F} \in q\alpha \cap (1 - q)\beta$

In this case, if we assume, as a result of sampling, that $(p - pq + q) = \frac{3}{4}$ and $(1 - p)(1 - q) = \frac{1}{4}$ we cannot determine the value of p and of q . But if p has first been ascertained with the help of THEOREM 16 and sampling then (at least when $p = q$) the result can be applied to THEOREM 17. For the derivation of this theorem we require T X, P 4, T XV, D 10, D 11, T 2, D 5, D 9, T 9, T 8, T 13, T XIV. The next next theorem is the theorem corresponding to THEOREM 17 in systems where there is no dominance.

THEOREM 18 $H\ 1.H\ 4b.H\ 5a.\supset.(\exists p)(\exists q).F_E(\alpha, \beta; \alpha, \beta) \in pqP \cap$
 $(p(1 - q) + q(1 - p))R \cap (1 - p)(1 - q)Q. G(\alpha, \beta, E) \cap$
 $\cap \mathcal{J} \in p\alpha \cap (1 - p)\beta. G(\alpha, \beta, E) \cap \mathcal{F} \in q\alpha \cap (1 - q)\beta$

In this case, if on the basis of sampling we assign a value to pq and to $(p(1 - q) + q(1 - p))$, then we can determine the values of p and q . The theorem requires: T X, P 4, T XV, D 10, D 11, T 2, D 5, T 9, T 8, T XIII, T 13.

Finally a theorem will be given which might have been known to Mendel. It is an example of a system which includes only *one* genetical unit. Suppose **F** and **M** are the females and males respectively of some species, suppose further that **g** and **h** are two mutually exclusive classes

of gametes and H a class of environments all satisfying the following conditions: (i) $\{F, g, H\}$ is a genetical unit; (ii) $D(g, h, H) \neq \Lambda$. $D(g, h, H) \subseteq M$. $G(g, h, H) \neq \Lambda$. $G(g, h, H) \subseteq g \cup h$. (iii) $D(h, h, H) = \Lambda$ (therefore $\{M, h, H\}$ is *not* a genetical unit); (iv) $S = \{F, M, g, h, H\}$ and S is *rand* $U \cap$ *rand* $D(H)$. If these conditions are satisfied we shall have:

$$F_H(g, g; g, h) \in \frac{1}{2}F \cap \frac{1}{2}M \text{ if and only if } G(g, h, H) \in \frac{1}{2}g \cap \frac{1}{2}h$$

THEOREM 19 $F \cap M = g \cap h = \Lambda$. $\{F, g, H\} \in$ *genunit*. $D(g, h, H) \neq \Lambda$.

$$D(g, h, H) \subseteq M. G(g, h, H) \neq \Lambda. G(g, h, H) \subseteq g \cup h. S = \\ = \{F, M, g, h, H\}. S \in \text{rand } U \cap \text{rand } D(H). \supset.$$

$$F_H(g, g; g, h) \in \frac{1}{2}F \cap \frac{1}{2}M. \equiv. G(g, h, H) \in \frac{1}{2}g \cap \frac{1}{2}h$$

This theorem requires for its derivation T 11, T XV, T VII, D 10, D 11, T XII, T 13.

We can now see clearly what was Mendel's discovery in the Christopher Columbus sense and what was his discovery in the J. J. Thomson sense distinguished above. His discovery in the first sense (inductive hypothesis) was the $\frac{3}{4}P \cap \frac{1}{4}Q$ frequencies in the offspring when hybrids are mated, if this is understood as being asserted (as above) for inaccessible sets. This is expressed in THEOREM 15. Mendel's discovery in the second sense (explanatory hypothesis) is the hypothesis that is expressed in H 4a. But we have seen that in this form it is unnecessary. The much weaker form of H 4c suffices, especially if we begin with T 16 and then, using its results with the value of p determined by sampling (coupled with the additional hypothesis: $p = q$), we pass to T 17. Where there is no dominance (in Mendel's experiments one phenotype is in each case dominant to the other) p and q can be determined independently of T 16 with the help of T 18. Thus the convenient minimum assumption is H 4b. It could be argued that the assumption of two kinds among the gametes of hybrids is not so much a discovery of the second kind as a special application of a general *causal principle* to embryology and genetics. But this does not mean that it cannot be discussed.

It is often said that Mendel discovered what is called particulate inheritance. But, except in the sense in which gametes are particles, Mendel did not specifically speak of particles. Strictly speaking a hypothesis involving cell parts only becomes important when we consider the

breakdown of Mendel's Second Law. The whole of Mendel's work can be expressed with the help of $D(\alpha, \beta, E)$, $G(\alpha, \beta, E)$ and $F_E(\alpha, \beta; \gamma, \delta)$ and thus in terms of gamete and environment classes, the classes of zygotes which can be formed with them and the classes of lives which develop from the zygotes in the environments.

The above analysis has shown the central role which is played by the hypotheses of random union of the gametes and of random development in obtaining the Mendelian ratios (see especially steps (2), (3) and (4) in the proof of Theorem 15). These do not receive the attention they deserve in genetical books. Sometimes they are not even mentioned. This is particularly true of the hypothesis of random development. That Mendel was aware of it is clear from the following passage in the translation from which we have already quoted (p. 340):

A perfect agreement in the numerical relations was, however, not to be expected, since in each fertilization, even in normal cases, some egg cells remain undeveloped and subsequently die, and many even of the well-formed seeds fail to germinate when sown.

In addition to the special hypotheses H 1 to H 5 there are also the six postulates to be taken into consideration. Any departure from these could affect the result. This provides plenty of scope for reflexion. But perhaps the most striking feature of the Mendelian systems is the fact that only one class of environments is involved and is usually not even mentioned. Some interesting discoveries may await the investigation of multi-environmental systems. Provision for this is made in Definitions 4, 7 and 11 and a variable having classes of environments as its values accompanies all the above biological functors. At the same time attention should be drawn to the fact that no provision is made, either here or in current practice, for mentioning the environments of the gametes. And yet it is not difficult to imagine situations in which the necessity for this might arise.

It will be noticed that no use has here been made of the words 'probability', 'chance', or 'independent', although these words are frequently used in genetical books with very inadequate explanation. Here the term 'random' has been used but its two uses have been explained in detail. In passing it may be mentioned that 'S is random with respect to F_E ' is also definable along analogous lines and then the Pearson-Hardy law is derivable.

In conclusion I should like to draw attention to the way in which the

foregoing analysis throws into relief the genius of Mendel, which enabled him to see his way so clearly through such a complicated situation. I also wish to express my thanks to Professor John Gregg of Duke University and to my son Mr Michael Woodger of the National Physical Laboratory for their help in the preparation of this article.

AXIOMATIZING A SCIENTIFIC SYSTEM BY AXIOMS IN THE FORM OF IDENTIFICATIONS

R. B. BRAITHWAITE

University of Cambridge, Cambridge, England

A scientific deductive system ("scientific theory") is a set of propositions in which each proposition is either one of a set of initial propositions (a "highest-level hypothesis") or a deduced proposition (a "lower-level hypothesis") which is deduced from the set of initial propositions according to logico-mathematical principles of deduction, and in which some (or all) of the propositions of the system are propositions exclusively about observable concepts (properties or relations) and are directly testable against experience. In this paper these testable propositions will be taken to be empirical generalizations of the form Every *A*-specimen is a *B*-specimen, whose empirical testability consists in the fact that such a proposition is to be rejected if an *A*-specimen which is not a *B*-specimen is observed. (Statistical generalizations of the form The probability of an *A*-specimen being a *B*-specimen is *p* can be brought within the treatment; here testability depends upon more sophisticated rejection rules in terms of the proportions of *B*-specimens in observed samples of *A*-specimens.) The object of constructing a scientific theory is to 'explain' empirical generalizations by deducing them from higher-level hypotheses.

A scientific deductive system will make use of a *basic logic* independent of the system to provide its principles of deduction. It will be convenient to assume that this basic logic includes all the deductive principles of the system, so that none of these are specific to the system itself and the deductive power of the system will be given by the addition to the basic logic of the system's set of initial propositions. The system can then be expressed by a formal axiomatic system (called here a *calculus*) in which the axioms (the "initial formulae") fall into two sets, one set consisting of those axioms required for the basic logic of the system (which set will be empty if the basic logic has no axioms) — no axiom of this set will contain any extra-logical constants — and another set of axioms containing non-vacuously extra-logical constants (Tarski's *proper axioms* [10, p. 306]) corresponding, one to one, to the set of initial propositions

of the scientific system. The rules of derivation of the calculus will then correspond to the deductive principles of the basic logic. Since we are not concerned with the nature of this basic logic we shall ignore the axioms and theorems of the calculus which forms a sub-calculus representing the basic logic and shall only be interested in *proper* axioms and *proper* theorems (i.e. those which contain non-vacuously extra-logical constants, which will be called *primitive terms*) and which are interpreted as representing the propositions of the scientific system. The theorems (or axioms) representing the directly testable propositions will be called *testable theorems* (or axioms), and the primitive terms occurring in these theorems *observable terms*.

The problem raised by scientific deductive systems for the philosophy (or logic or semantics) of science is to understand how the calculus is interpreted as expressing the system. If all the proper axioms are testable axioms, and consequently all the proper theorems are testable theorems, there is no difficulty, since all the extra-logical terms (i.e. primitive terms) occurring in the calculus are observable terms so that all the proper axioms and theorems can be interpreted as propositions directly testable by experience. The semantic rules for the interpretation of the calculus by means of direct testability apply equally to all the proper axioms and theorems; so the calculus can be interpreted all in a piece.

But the situation is different for the deductive system of a more advanced science which makes use in its initial propositions of concepts (call them *theoretical concepts*) which are not directly observable, so that the propositions containing these are not directly testable. Here the axioms of the calculus contain primitive terms which are not observable terms, and these *theoretical terms* have to be given an interpretation not by a semantic rule concerning direct testability but by the fact that testable theorems are derivable from them in the calculus. The calculus is thus interpreted from the bottom upwards: the testable theorems are interpreted by a semantic rule of direct testability, and the other theorems and axioms are then interpreted syntactic-osemantically by their syntactic relations to the testable theorems. Theoretical terms are not definable by means of observable concepts — the 'reductionist' programme of thorough-going logical constructionists and operationalists cannot profitably be applied to the theoretical concepts of a science — though they may be said to be *implicitly* defined by virtue of their place in a calculus which contains testable theorems. The empirical interpretation of the calculus is thus given by a directly empirical interpretation of the testable axioms

and theorems and an indirectly empirical interpretation of the remainder. (For all this see R. B. Braithwaite [3, Chapter III].)

In order that a calculus containing theoretical terms should be able to be interpreted in this indirectly empirical way, it is necessary that each of the observable terms should occur in at least one of the proper axioms. These may be divided into three categories: (1) Testable axioms whose primitive terms are all observable terms; (2) Axioms whose primitive terms are all theoretical terms: these will be called *Campbellian axioms*, since collectively they represent N. R. Campbell's "hypothesis" consisting of "statements about some collection of ideas which are characteristic of the [scientific] theory" ([4], p. 122), and the highest-level hypotheses represented by them will be called *Campbellian hypotheses*; (3) Axioms whose primitive terms are both observable terms and theoretical terms: these will be called *dictionary axioms*, since they correspond to Campbell's "dictionary". Since no philosophical problems arise in connexion with testable axioms, we will suppose that there are no testable axioms in the calculus, so that no direct empirical interpretation is possible at the axiom level. To simplify our discussion we will further suppose that each dictionary axiom is of the form of an identity

$$a = (\dots \lambda_1 \dots \lambda_2 \dots)$$

where a is an observable term standing alone on the left-hand side of the identity with the right-hand side containing only theoretical terms λ_1 , λ_2 , etc. as primitive terms. Dictionary axioms in this form will be called *identificatory axioms*, since they may be said to 'identify' an observable term by means of theoretical terms. In order that these identificatory axioms should be able to function in a calculus to be interpreted as a scientific system, the basic logic governing the identity sign will be assumed to be strong enough to permit the derivation from an axiom of the form $a = (\dots \lambda_1 \dots \lambda_2 \dots)$ of every theorem obtained by substituting a for $(\dots \lambda_1 \dots \lambda_2 \dots)$ at any place in any axiom or theorem in which $(\dots \lambda_1 \dots \lambda_2 \dots)$ occurs.

The simplified calculi to be considered will thus contain, as proper axioms, Campbellian axioms concerned with the theoretical terms of the scientific calculus and identificatory axioms relating the observable terms of the calculus to the theoretical terms by identifying each of the former with a logical function of the latter. If a is an n -ary predicate, an identificatory axiom $a = (\dots \lambda_1 \dots \lambda_2 \dots)$ will, with a suitable basic logic,

permit the derivation from this axiom of

$$(x_1)(x_2)\dots(x_n)(a(x_1, x_2, \dots, x_n) \equiv Q(x_1, x_2, \dots, x_n)),$$

where Q is an abbreviation for $(\dots\lambda_1\lambda_2\dots)$, so that a will be *definable* with respect to the identificatory axiom (together with the basic logic) in terms of λ_1, λ_2 , etc. in E. W. Beth's sense of "definable" [(2), p 335]. (In [3, p. 57] I called sentences of the form $a = (\dots\lambda_1\lambda_2\dots)$ *definitory formulae*; but I now prefer to call them identificatory axioms (or theorems) and to reserve the word "definition" and its cognates for notions which are semantical and not purely syntactical.)

Most axiomatizations of a scientific theory contain Campbellian axioms among their proper axioms. Philosophers of science frequently think that it is the Campbellian axioms representing the Campbellian hypotheses which express the essence of the theory, the dictionary axioms (which in the simplest cases are identificatory axioms) having the function of 'semantical rules' or 'co-ordinating definitions' or 'definitory stipulations' relating the observable terms to the theoretical terms. Thus there would be an absolute distinction between Campbellian and dictionary axioms. It would follow from this point of view that a calculus which makes use of theoretical terms must include Campbellian axioms if it is to be interpreted to express what is of importance in the scientific theory. This, however, is not the case. Calculi whose proper axioms are all identificatory can serve to express empirical deductive systems: indeed, given a calculus which contains Campbellian axioms, it is sometimes possible to construct another calculus having the same theoretical terms whose proper axioms are all identificatory which is *testably equivalent* to the first calculus in the sense that the testable theorems of the two calculi are exactly the same.

This will always be the case if the basic logic of the calculus is simple enough. We will consider the case in which the basic logic is merely that of propositional logic combined with that of the first-order monadic predicate calculus with identity (and with a finite number of predicates). This basic logic is also that of finite Boolean lattices, and it will be convenient to regard it as expressed by a calculus (called a *Boolean calculus*) whose logical constants are, besides those of the propositional calculus, constants whose class interpretation is union (\cup), intersection (\cap), complementation ($'$), the universe class (e), the null class (o), class inclusion (\subset) and class identity ($=$). This basic logic is sufficient for the construction of theories in which empirical generalizations of the form Every AB ...specimen is a K -specimen (represented in the calculus by a testable theorem

$(a \cap b \cap \dots) \subset k$, a being interpreted as designating the class of A -specimens, and similarly for the other small italic letters) are explained as deducible from initial propositions containing theoretical class-concepts designated by $\lambda_1, \lambda_2, \dots$ (Simple examples of such theories are given in [3, Chapters III and IV].) Since all the propositions concerned will be universal propositions (i.e., of the form Every...-specimen is a ...-specimen), every formula of the calculus is equivalent to a formula in normal form $\dots = o$.

Let \mathfrak{S}_1 be a calculus of this type comprising n identificatory axioms $D_1, D_2, \dots D_n$ identifying the n observable terms $a_1, a_2, \dots a_n$ by means of l theoretical terms $\lambda_1, \lambda_2, \dots \lambda_l$.

D_r is $a_r = \Delta_r$, where Δ_r is a Boolean expression whose terms are all theoretical terms. Let the calculus also comprise m Campbellian axioms $C_1, C_2, \dots C_m$ containing theoretical terms alone. Derive from C_r the equivalent formula $I_r = e$, where I_r is a Boolean expression whose terms are all theoretical terms, and let Γ be $(\Gamma_1 \cap \Gamma_2 \cap \dots \Gamma_m)$.

Now consider a related calculus \mathfrak{S}_2 containing the same observable and theoretical terms but with no Campbellian axioms. Let its n identificatory axioms be $E_1, E_2, \dots E_n$, where E_r is $a_r = (\Delta_r \cap \Gamma)$. We will prove that (under a weak condition) \mathfrak{S}_2 is testably equivalent to \mathfrak{S}_1 in that the testable theorems in each calculus, obtained in each case by eliminating the theoretical terms from the axioms, are the same.

The proof depends upon the classical theory of elimination of variables from Boolean equations and is a development of a result of A. N. Whitehead [11, p. 60, (5) and p. 65, (1)]. Consider the 'universe' of the l theoretical terms $\lambda_1, \lambda_2, \dots \lambda_l$ (these are common to both \mathfrak{S}_1 and \mathfrak{S}_2). The 2^l *minimals* $(\lambda_1 \cap \lambda_2 \cap \dots \lambda_l), (\lambda_1 \cap \lambda_2 \cap \dots \lambda_l'), \dots (\lambda_1' \cap \lambda_2' \cap \dots \lambda_l')$ form a partition of the universe (in accordance with the basic logic of finite Boolean lattices), i.e. using a suffixed μ to designate a minimal, $\mu_r \cap \mu_s = o$ for $r \neq s$; $\bigcup_i \mu_i = e$. Then Δ_r , the Boolean expression whose

terms are all theoretical terms which is identified with a_r by the identificatory axiom D_r of \mathfrak{S}_1 , is the union of the minimals in some sub-set of the minimals; and D_r is equivalent to $a_r = \bigcup_{i: \mu_i \subset \Delta_r} \mu_i$ and, in normal form, to

$$(a_r' \cap \bigcup_{i: \mu_i \subset \Delta_r} \mu_i) \cup (a_r \cap \bigcup_{i: \mu_i \subset \Delta_r'} \mu_i) = o.$$

If D is $D_1.D_2.\dots D_n$, D is then equivalent to $\bigcup_i (A_i \cap \mu_i) = o$, where

$$A_i \text{ is } \bigcup_{j: \mu_j \subset \Delta_i} a_j' \cup \bigcup_{j: \mu_j \subset \Delta_i'} a_j).$$

If C is $C_1.C_2\dots C_m$ (the conjunction of the Campbellian axioms), C is equivalent to $\bigcup_{j:\mu_i \subset \Gamma'} \mu_i = o$. $C.D$ is then equivalent to

$$\bigcup_{i:\mu_i \subset \Gamma} (A_i \cap \mu_i) \cup \bigcup_{i:\mu_i \subset \Gamma'} (e \cap \mu_i) = o.$$

The resultant in normal form R_1 of eliminating all the minimals from $C.D$ is $\bigcap_{i:\mu_i \subset \Gamma} A_i = o$. R_1 is equivalent to the conjunction of all the testable theorems; so a testable formula T is a theorem of \mathfrak{S}_1 if and only if $R_1 \supset T$.

By a similar argument applied to the axioms of \mathfrak{S}_2 , E_r is equivalent to $a_r = \bigcup_{i:\mu_i \subset (\Delta_r \cap \Gamma)} \mu_i$ and, in normal form, to

$$(a_r' \cap \bigcup_{i:\mu_i \subset (\Delta_r \cap \Gamma)} \mu_i) \cup (a_r \cap \bigcup_{i:\mu_i \subset (\Delta_r' \cap \Gamma)} \mu_i) \cup (a_r \cap \bigcup_{i:\mu_i \subset \Gamma'} \mu_i) = o.$$

If E is $E_1.E_2\dots E_n$, E is then equivalent to $\bigcup_i (B_i \cap \mu_i) = o$, where,

for an i such that $\mu_i \subset \Gamma$, B_i is $(\bigcup_{j:\mu_i \subset \Delta_j} a_j' \cup \bigcup_{j:\mu_i \subset \Delta_j'} a_j)$;

for an i such that $\mu_i \subset \Gamma'$, B_i is $\bigcup_{j:\mu_i \subset \Delta_j'} a_j$.

The resultant in normal form R_2 of eliminating all the minimals from E is $\bigcap_i B_i = o$, which, since $B_i = A_i$ for every i such that $\mu_i \subset \Gamma'$, is equivalent to $\bigcap_{i:\mu_i \subset \Gamma} A_i \cup \bigcup_j a_j = o$. A testable formula T is a theorem of \mathfrak{S}_2 if and only if $R_2 \supset T$.

Now impose the weak condition that Γ should not be wholly included within $\bigcup_j \Delta_j$, i.e. $\Gamma \neq (\Gamma \cap \bigcup_j \Delta_j)$. Under this condition there is at least one minimal, say μ_s , which is such that both $\mu_s \subset \Gamma$ and $\mu_s \subset \Delta_j'$ for every j . Then for this s , $A_s = B_s = \bigcup_j a_j$; and R_2 , like R_1 , is

$\bigcap_{i:\mu_i \subset \Gamma} A_i = o$. Hence $R_1 \equiv R_2$; and T is a testable theorem of \mathfrak{S}_1 if and only if T is a testable theorem of \mathfrak{S}_2 .

Thus, unless the Campbellian axioms C of \mathfrak{S}_1 restrict the universe of theoretical terms to a class Γ which is included in the union of all the observable terms according to their identifications in \mathfrak{S}_1 , the calculus \mathfrak{S}_2 , constructed from \mathfrak{S}_1 by omitting its Campbellian axioms and substituting $(\Delta_r \cap \Gamma')$ for Δ_r in each of its identificatory axioms, is testably equivalent to \mathfrak{S}_1 in the sense that every testable theorem of the one is also a testable theorem of the other.

In Whitehead's language [11, p. 59] each identificatory axiom is *unlimiting* with respect to all the theoretical terms simultaneously in the sense that the resultant of eliminating the observable term from the axiom is equivalent to $o = o$, a theorem of the basic logic. A calculus such as \mathfrak{S}_2 whose proper axioms are all identificatory therefore imposes no limitation upon the universe (Whitehead's *field*) of the theoretical terms. Such a limitation is imposed by the Campbellian axioms of \mathfrak{S}_1 . But, if this limitation restricts the theoretical-term universe to a universe which falls wholly within the class which is the union of all the observable terms, it will be impossible in the future to adapt the scientific theory expressed by a calculus using theoretical terms limited in this way to explain new empirical generalizations relating some of the observable concepts to new observable concepts not concerned in the original theory [3, pp. 73ff.]. An axiomatization of a scientific theory which is capable of being adapted in this way must not impose such a drastic limitation upon its theoretical terms. So our result may be put in the form that to every adaptable calculus comprising Campbellian axioms a testably equivalent calculus can be constructed all of whose proper axioms are identificatory.

This result has been established only for a scientific system which makes use of a very simple basic logic (that of finite Boolean lattices); and the extent to which it can be generalized to apply to systems comprising Campbellian hypotheses and using more powerful basic logics requires investigation. That it is possible to have a theory using a mathematical basic logic in whose calculus theorems are derived from identificatory axioms alone is shown by such a simple example as that of explaining $a^2 + b^2 = 1$, where a and b stand for observably determined numbers, by identifying a with $\sin \theta$ and b with $\cos \theta$, θ being a theoretical 'parameter'.

One obvious qualification must be made. If identificatory axioms in the form of a description $a = (\iota x)(\phi(x))$ are permitted, and if their underlying logic is similar to Russell's doctrine of descriptions in that $(\exists x)(\phi(x))$ is derivable from any formula containing $(\iota x)(\phi(x))$, an identificatory axiom for a calculus \mathfrak{S}_3 of the form $a_r = (\iota x)(x = \Delta_r. \Gamma = e)$ would imply both $a_r = \Delta_r$ and $\Gamma = e$, and all the axioms of \mathfrak{S}_1 would be derivable from those of \mathfrak{S}_3 , a stronger system. But every theoretical scientist would regard the proposal to substitute a theory expressed by \mathfrak{S}_3 for one expressed by \mathfrak{S}_1 as a logician's trick. So for scientific discussion the notion of identificatory axiom must be restricted to one from which alone no Campbellian theorem can be derived, i.e. an identificatory axiom must be unlimiting with respect to all its theoretical terms simultaneously.

The possibility, in suitable cases, of constructing a testably equivalent calculus comprising only identificatory proper axioms is very relevant to the discussion among philosophers of science as to whether or not some of the highest-level hypotheses of the scientific theory expressed by the calculus should be regarded as analytic or logically necessary rather than as factual or contingent. It is admitted by all empiricists that the conjunction of all the hypotheses must be contingent, since together they have empirically testable consequences. But, if the highest-level hypotheses contain theoretical concepts, it is never from one of these hypotheses alone but always from a conjunction of them that testable propositions are deducible; and so the possibility is left open that some of these hypotheses are not contingent, and hypotheses representing dictionary axioms (e.g. the identificatory axioms considered in this paper) are frequently held to be analytic. For example, A. J. Ayer [1, p. 13], in his account of the "indirect verifiability" of scientific statements (which is similar to mine), explicitly allows that the conjunctions whose consequences are "directly verifiable" may include analytic statements, his reason being that "while the statements that contain [theoretical] terms may not appear to describe anything that anyone could ever observe, a 'dictionary' may be provided by means of which they can be transformed into statements that are verifiable; and the statements which constitute the dictionary can be regarded as analytic". And E. Nagel [9, pp. 209f.], in a recent discussion of my book [3], criticises me for my "disinclination to regard as 'absolute' Norman Campbell's distinction between the 'hypotheses' and the 'dictionary' of a theory. In Campbell's analysis, the hypotheses postulate just what relations hold between the purely theoretical but otherwise unspecified terms of a theory, while the dictionary provides the co-ordinating definitions for some of the theoretical terms or for certain functions of them". "Every testable theory must include a sufficient number of co-ordinating definitions which are not subject to experimental control"; and, though Nagel never explicitly says that co-ordinating definitions state analytic propositions, he declares that they have "the status of semantic rules" and contrasts them with "factually testable assumptions" and with "genuine hypotheses". The existence of calculi with no Campbellian axioms representing "genuine hypotheses" and the possibility in suitable cases of converting calculi having Campbellian axioms into calculi with only identificatory proper axioms make it impossible to ascribe a logically necessary status to what is represented by the identificatory axioms taken all together. Since it is the whole set

of the hypotheses that conjunctively are "subject to experimental control", it is possible that some sub-set of them are not so subject. But there would seem to be no good reason for placing any of the identificatory axioms in this latter category. Nagel goes so far as to say that in the simplest calculus which I gave as an example [3, pp. 54ff], in which $a = (\lambda \cap \mu)$, $b = (\mu \cap \nu)$, $c = (\nu \cap \lambda)$ are the axioms and $(a \cap b) \subset c$, $(b \cap c) \subset a$, $(c \cap a) \subset b$ are the testable theorems, "the obvious (and I think correct) alternative to Braithwaite's account is to construe two of the equational formulas in the [axiom set] not as hypotheses but as having the function of semantical *rules* . . . which *assign* partial meanings to the theoretical terms and to count the remaining formula as a genuine hypothesis when such definitory stipulations have once been laid down." But he gives no way of selecting the one "genuine hypothesis" from among the three which appear in the complete symmetry.

There would seem to be a stronger case for regarding Campbellian hypotheses as logically necessary and for accounting for the contingency of the lowest-level generalizations by the contingency of the identifications provided by identificatory axioms. The function of Campbellian axioms is always that of limiting the universe of the theoretical terms, left unlimited by the identificatory axioms; and it can be said that, since the theoretical scientist in constructing a theory to explain his empirical generalizations has great liberty of choice in selecting his theoretical terms, he may well in the act of selecting them impose a limitation upon the 'degrees of freedom' of their universe by a set of Campbellian axioms, and this limitation (i.e. the conjunction of the Campbellian axioms) will never by itself be "subject to experimental control", since it is concerned only with theoretical terms. But, in a calculus comprising both identificatory and Campbellian axioms, the testable theorems derivable from the former axioms form only a sub-class of the testable theorems derivable from the conjunction of all the axioms; so the Campbellian axioms may be given an empirical interpretation by virtue of the testability of the additional theorems which are derivable by adding them to the identificatory axioms. There is no adequate reason for refusing to interpret every proper axiom in a calculus expressing a scientific theory as representing a contingent proposition, the empirical interpretation of the axioms being given by the syntactical relations of the whole set of axioms to the testable theorems derivable from them. The only exception would be the uninteresting case in which a redundant theoretical term is introduced into a calculus by an axiom identifying it with a logico-mathematical

function of other theoretical terms. Such a *sterile* axiom ([3], p. 113), functioning merely as an abbreviatory device, may rightly be regarded as analytic.

There is one other consideration which has tended to confuse the issue in the minds of some scientists and philosophers of science. If, as is usually the case, the basic logic of the scientific theory is expressible by a calculus with axioms and theorems interpreted as propositions of logic or mathematics, these theorems will contain no extra-logical constants; and the use of one of these theorems in the derivation of a proper theorem of the scientific calculus will require an intermediate step in which a logical theorem is *applied* to the primitive terms concerned. If the logical sub-calculus uses the device of variables, this application will be effected by making substitutions of primitive terms for some or all of these variables. The theorem so derived will not be a proper theorem of the calculus, since the primitive terms will occur in it only vacuously, but neither will it be a theorem of the logical sub-calculus since it will contain primitive terms as extra-logical constants. Call such a theorem an *applicational theorem*. (For example, the derivation of $(a \cap b) \subset c$ from $a = (\lambda \cap \mu)$, $b = (\mu \cap \nu)$, $c = (\nu \cap \lambda)$ will require (if the basic logic is expressed as a Boolean calculus) the use of the applicational theorem $(\mu \cap \mu) = \mu$, which is not itself a theorem of a Boolean calculus but is derived from the Boolean theorem (or axiom) $(x \cap x) = x$, where x is a free variable with class symbols as substitution values.)

Applicational theorems fall in the no-man's-land between the theorems of the basic-logic sub-calculus and the proper theorems of the calculus. If the scientific part of the whole calculus is regarded not (as we have thought of it) as consisting of proper axioms and theorems (i.e. those containing primitive terms non-vacuously), but as consisting of all the axioms and theorems which are not comprised in the basic-logic sub-calculus (i.e. those which contain primitive terms either vacuously or non-vacuously), then the applicational theorems will be classed as falling within the scientific part. Since they will usually function there as premisses from which, together with the proper axioms, proper theorems are derived, and will not themselves be derived within this scientific part, it will be natural to class them, within this scientific part, with the proper axioms rather than with the proper theorems. A person who takes this point of view will then hold that the scientific part of the calculus comprises axioms which are to be interpreted as representing logically necessary propositions, these 'pseudo-axioms' being applicational theo-

terms whose interpretations are logically necessary by virtue of being applications to the concepts concerned of the laws of logic or mathematics. When, as is usually the case, the primitive terms concerned in the applicational pseudo-axioms are theoretical terms, these pseudo-axioms will simulate Campbellian axioms; and if the calculus is one all of whose proper axioms are identificatory, it will be described by a person who mistakes such applicational pseudo-axioms for Campbellian axioms not as a calculus with no Campbellian axioms, but as a calculus whose Campbellian axioms represent Campbellian hypotheses which are logically necessary. If this person also does not regard identificatory axioms as representing "genuine hypotheses", he may well assert that all the "genuine hypotheses" of the theory expressed by the calculus are logically necessary or *a priori*.

In our own time the thesis that the fundamental laws of physics are *a priori* has been maintained by A. S. Eddington, who has attempted to infer them, including the pure numbers which occur in them, from "epistemological considerations" ([7]), p. 57). The reasons Eddington gave at different places in his writings for his general thesis are different and doubtfully consistent, but his principal reason would seem to be an argument on Kantian lines that "the fundamental laws and constants of physics. . . are a consequence of the conceptual frame of thought into which our observational knowledge is forced by our method of formulating it, and can be discovered *a priori* by scrutinising the frame of thought" ([7], p. 104). Such a view is incompatible with an empiricist philosophy of science. But Eddington's programme of constructing a unified theory for physics whose fundamental hypotheses are to be *a priori* appears in a new light if his goal is described negatively as a theory having no contingent Campbellian hypotheses. For his goal would then, on our way of thinking, be a theory with no Campbellian hypotheses at all, represented by a calculus whose proper axioms were all identificatory; and we should explain his attribution of apriority to such a theory by his having mistaken for Campbellian axioms the applicational pseudo-axioms required to apply the basic logic to the concepts of the theory. And a programme of constructing a Campbellian-hypothesis-free system of physics, unhelpful though it may appear to a physicist, is not ridiculous to an empiricist philosopher.

Perhaps because Eddington was not interested in axiomatics this way of looking at his programme never, it seems, occurred to him. But scattered throughout his writings (e.g. [6], pp. 3, 242; [7], pp. 41, 134;

[8], p. 265) are many references to the essential part to be played by "identification" and "definition" in relating observation to theory, and he does not suppose that these identifications are *a priori*: "we cannot foresee what will be the correspondence between elements in [the] *a priori* physical description and elements in our familiar apprehension of the universe" [7, p. 134]. That Eddington's ideal was a system with no Campbellian axioms is suggested by his preferring the theory of numbers to geometry as an analogue for a system of physics: "If the analogy with geometry were to hold good, there would be a limit to the elimination of hypothesis, for a geometry without any axioms at all is unthinkable. But... [in] the theory of numbers... there is nothing that can be called an axiom. We shall find reason to believe that this is in closer analogy with the system of fundamental laws of physics" [7, p. 45]. So I think it is a fair, and charitable, gloss on Eddington to take his programme as the constructing for the whole of physical theory of an *identificatory system*, whose axiomatization would comprise only identificatory proper axioms, in contrast with the programme of all other theoretical physicists of constructing *Campbellian systems*, whose axiomatization would comprise Campbellian as well as identificatory axioms.

Can anything in general be said as to the relative advantages of constructing Campbellian or identificatory systems as explanatory scientific theories? Not much more, I think, than that, since the calculus expressing a Campbellian system will be stronger (by virtue of comprising Campbellian axioms and theorems) than a testably equivalent identificatory calculus (in which no Campbellian theorem can be derived), a Campbellian system can probably be more easily adapted in the future to explain new empirical generalizations, as is illustrated in the history of physics by the great adaptability of systems which included the conservation of energy as a Campbellian hypothesis. An identificatory system would seem to be the more appropriate one for providing the most economical theory to explain a closed set of empirical generalizations. But it may well be the case that there are subjects, perhaps those of some of the social sciences, in which identificatory systems are those which arise most naturally in reflecting upon the subject-matter concerned. The development of the social sciences has been retarded by a false belief that numerical mathematics provides the only deductive techniques so that, to construct a scientific theory, it is necessary that both the observable and the theoretical concepts of a science should be numerically measurable. It may also have been retarded by a false belief that a science can only use

theoretical concepts if these can be related together in Campbellian hypotheses. A realization by social scientists that there is no need to imitate the methods of theory-construction which have proved so successful in the physical sciences, and that theories whose theoretical concepts occur only in hypotheses 'identifying' the observable concepts are perfectly good explanatory theories (provided, of course, that testable consequences can be deduced from these hypotheses), might encourage them to a greater boldness in thinking up theoretical concepts and trying out theories containing them. This sort of encouragement is the contribution a philosopher of science can make the progress of science.

One last and philosophical remark. To identify, by means of an identificatory axiom, an observable term with a logico-mathematical function of theoretical terms in a calculus expressing a scientific theory is one way of *explicating* (in the sense of R. Carnap [5, Chapter I]) the "inexact concept" for which the observable term stands in ordinary language. To propose a scientific theory containing theoretical concepts which is to be testable against experience involving inexact concepts requires explications of these concepts; and, if the theory is an identificatory system, the hypotheses of the theory will consist entirely of such explications. Conversely, a set of explications by means of theoretical concepts will constitute the hypotheses of an identificatory system; and, if this system permits the deduction of empirically testable consequences, it will be a scientific theory. A philosopher propounding such a system of explications must not be dismissed as a rationalist metaphysician on the sole ground that the hypotheses of his system appear all in the form of new 'definitions'. His system will only fail to be scientific if nothing empirical follows from all his definitions taken together.

Bibliography

- [1] AYER, A. J., *Language Truth and Logic* (2nd edition). London 1946, 160 pp.
- [2] BETH, E. W., *On Padoa's method in the theory of definition*. *Indagationes Mathematicae*, vol. 15 (1953), pp. 330-339.
- [3] BRAITHWAITE, R. B., *Scientific Explanation*. Cambridge 1953, X + 376 pp.
- [4] CAMPBELL, N. R., *Physics The Elements*. Cambridge 1920, X + 565 pp.
- [5] CARNAP, R., *Logical Foundations of Probability*. Chicago 1950, XVIII + 607 pp.
- [6] EDDINGTON, (Sir) A. S., *Relativity Theory of Protons and Electrons*. Cambridge 1936, VIII + 336 pp.

- [7] —, *The Philosophy of Physical Science*. Cambridge 1939, X + 230 pp.
- [8] —, *Fundamental Theory*. Cambridge 1946, VIII + 292 pp.
- [9] NAGEL, E., *A budget of problems in the philosophy of science*. The Philosophical Review, vol. 66 (1957), pp. 205–225.
- [10] TARSKI, A., *Some methodological investigations on the definability of concepts*. Chapter X in *Logic, Semantics, Metamathematics*. Oxford 1956, XIV + 471 pp.
- [11] WHITEHEAD, A. N., *Universal Algebra*. vol. 1, Cambridge 1898, XXVI + 586 pp.

DEFINABLE TERMS AND PRIMITIVES IN AXIOM SYSTEMS

HERBERT A. SIMON

Carnegie Institute of Technology Pittsburgh, Pennsylvania, U.S.A.

An axiom system may be constructed for a theory of empirical phenomena with any of a number of goals in mind. Some of these goals are identical with those that motivate the axiomatization of mathematical theories, hence relate only to the formal structure of the theory — its syntax. Other goals for axiomatizing scientific theories relate to the problems of verifying the theories empirically, hence incorporate semantic considerations.

An axiom system includes, on the one hand, entities like primitive terms, defined terms, and definitions, and on the other hand, entities like axioms, theorems, and proofs. Tarski [10, p. 296] has emphasized the parallelism between the first triplet of terms and the second. The usual goals for axiomatizing deductive systems are to insure that neither more nor less is posited by way of primitive terms and axioms than is necessary and sufficient for the formal correctness of the definitions and proofs, and hence the derivability of the defined terms and theorems. An axiom system is usually accompanied by proofs of the independence, consistency, and completeness of its axioms; and presumably should also be accompanied — although it less often is — by proofs of the independence, consistency, and completeness of its primitive terms.

Frequently a set of sentences (axioms and theorems) and terms admits alternative equivalent axiom systems: that is non-identical partitionings of the sentences into axioms and theorems, respectively; and of the terms into primitive and defined terms. Hence, a particular set of axioms and primitive terms may be thought of as a (not necessarily unique) basis for a class of equivalent axiom systems.

In constructing an axiom system for an empirical theory, we may wish to distinguish sentences that can be confronted more or less directly with evidence (e.g., “the temperature of this water is 104°”) from other sentences. We may wish to make a similar distinction between predicates, functors, and other terms that appear in such sentences (e.g., “temperature”) and those that do not. The terms “observation sentences” and

"observables" are often used to refer to such sentences and such terms, respectively.¹

The distinction between observables and non-observables is useful in determining how fully the sentences of a theory can be confirmed or disconfirmed by empirical evidence, and to what extent the terms of the theory are operationally defined. In addition to the formal requirements, discussed previously, we might wish to impose the following additional conditions on an axiom system for an empirical theory:

(1) that the entire system be factorable into a subsystem that is equivalent to some axiom system for a part of logic and mathematics, and a remainder;

(2) that in the remainder, axioms correspond to observation sentences, and primitive terms to observables.

Condition (2) is, of course, a semantic rather than a syntactic condition, and has no counterpart in the axiomatization of mathematical theories. The usefulness of the condition is that, if it is met, the empirical testability of observation sentences guarantees the testability of all the sentences in the system, and the operational definability of observables guarantees the operationality of all the terms. In the remainder of this paper we shall explore some problems that arise in trying to satisfy Condition (2), and some modifications in the notion of definability — as that term is used in formal systems — that are needed to solve these problems.

The question of what characteristics an axiom system should possess has been raised in the past few years [9] in connection with the definability of mass in Newtonian mechanics. In one recent axiomatization of Newtonian particle mechanics [5] particular care is taken to meet the syntactic conditions for a satisfactory axiomatization, and mass is introduced as a primitive term. In another axiomatization [8] special attention is paid to semantic questions, and definitory equations for mass are introduced.

Definability and Generic Definability. Tarski [10] has proposed a definition of the term *definability* in a deductive system, and has shown how this definition provides a theoretical foundation for the method employed by Padoa [6] to establish whether particular terms in a system are definable or primitive. In their axiomatization of classical particle mechanics, McKinsey, Sugar and Suppes [5, Paragraph 5] employ the method of Padoa to show that, by Tarski's definition, mass and force are

¹ For a more extended discussion of these terms, see [2, pp. 454–456].

primitive terms in their system. Application of the same method to Simon's earlier axiomatization of Newtonian mechanics [8] gives the same result — mass and force are primitives in that system.

The latter result appears to conflict with common-sense notions of definability, since in [8] the masses of the particles can (in general) be computed when their positions and accelerations are known at several points in time [8, Theorem I]. Condition (2) of the previous section is violated if masses, which are not observables, are taken as primitive terms; and it appears paradoxical that it should be possible to calculate the masses when they are neither observables nor defined terms. These difficulties suggest that Tarski's concept of definability is not the most satisfactory one to use in the axiomatization of empirical science.

A closer examination of the situation, for [8], shows that the masses are not uniquely determined in certain situations that are best regarded as special cases — e.g., the case of a single unaccelerated particle. It is by the construction of such special cases, and the application of the method of Padoa to them, that McKinsey, Sugar and Suppes show mass to be a primitive in [5], and by inference in [8]. But I shall show that if the definition of Tarski is weakened in an appropriate way to eliminate these special cases it no longer provides a justification for the method of Padoa, but does provide a better explication of the common-sense notion of definability.

Statement of the Problem. We shall discuss the problem here in an informal manner. The treatment can easily be formalized along the lines of Tarski's paper.² In Tarski's terms [10, p. 299], *the formula $\phi(x; b', b'', \dots)$ defines the extra-logical constant a if, for every x , x satisfies ϕ if and only if x is identical with a ; i.e., if:*

$$(I) \quad (x):x = a. \equiv .\phi(x; b', b'', \dots),$$

where x is the only real variable in ϕ , and b', b'', \dots are the members of a set of extra-logical constants (primitives and/or defined terms).

Translated into these terms, the (attempted) definition of "the mass of particle i " in [8, p. 892] proceeds thus: (1) We take as the function ϕ the conjunction of the six scalar equations that state the laws of conservation of momentum and conservation of angular momentum for a system of particles. (2) We take as the set B the paths of the particles in some time

² Compare also [2, p. 439].

interval. (3) We take as x the set of numbers m_i , that satisfy ϕ for the given B .

This procedure does not satisfy Tarski's definition since the existence and uniqueness of the masses is not guaranteed. For example, in the case of a single, unaccelerated particle, *any* number, m , substituted in the equations for conservation of momentum and angular momentum will satisfy those equations. But Tarski shows (his Theorem 2) that if two constants satisfy a definitory formula for a particular set, B , they must be identical.

Generic Definition. To remove the difficulty, we replace Tarski's definition with a weaker one: *the formula $\phi(x; b', b'', \dots)$ DEFINES GENERICALLY the extralogical constant a if, for every x , if x is identical with a , x satisfies ϕ :*

$$(I') \quad (x): x = a \supset \phi(x; b', b'', \dots).$$

After the equivalence symbol in formula (I) has been replaced by an implication in this way, the three theorems of Tarski's paper are no longer provable. In particular, formula (7) in his proof of Theorem I [10, pp. 301-302] can no longer be derived from the modified forms of his formulas (3) and (6). Hence, the method of Padoa cannot be used to disqualify a proposed generic definition.

It is easy to show that in [8] mass is generically defined by means of the paths of the particles on the basis of the Third Law of Motion (more exactly, the laws of conservation of momentum and angular momentum); and that resultant force is generically defined by means of the paths of the particles and their masses on the basis of the Third and Second Laws of Motion [8, p. 901]. Similarly, we can show that in [5, p. 258] resultant force is generically defined by means of the paths of the particles and their masses on the basis of the Second Law of Motion.

The advantage of substituting generic definition for definition is that, often, a constant is not uniquely determined for all possible values of the other extra-logical constants, but experimental or observational circumstances can be devised that do guarantee *for those circumstances* the unique determination of the constant.

In the axiom system of [8], for example, the conditions under which masses exist for a system of particles and the conditions under which these masses are unique have reasonable physical interpretations. The observables are the space-time coordinates of the particles. From a

physical standpoint, we would expect masses (not necessarily unique) to be calculable from the motion of a set of particles, using the principles of conservation of momentum and angular momentum, whenever this set of particles was physically isolated from other particles. Moreover, we would expect the relative masses to be uniquely determined whenever there was no proper subset of particles that was physically isolated from the rest. These are precisely the conditions for existence (Definition 3) and uniqueness (Theorem I and Definition 6) of the masses in this axiomatization. Thus, the definition of mass in [8] does not lead to a unique determination of the mass of a single star at a great distance from other stars, but does permit the calculation, uniquely up to a factor of proportionality, of the masses of the members of the solar system from observation of their paths alone, and without postulating a particular force law [8, pp. 900–901].

OTHER CONCEPTS OF DEFINABILITY

The sharp distinctions between axioms and theorems, and between primitive and defined terms have proved useful dichotomies in axiomatizing deductive systems. We have seen that difficulties arise in preserving the latter distinction in empirical systems, when the axiom system is required to meet Condition (2) — when primitive terms are identified with observables. But it has long been recognized that comparable difficulties arise from the other half of Condition (2), that is, from the identification of axioms with observation sentences. In our axiomatization of Newtonian mechanics, for example, the law of conservation of momentum, applied to an isolated system of particles, is an identity in time containing only a finite number of parameters (the masses). If time is assumed to be a continuous variable, this law comprises a nondenumerable infinity of observation sentences. Hence, the law is not itself an observation sentence nor is it derivable from a finite set of observation sentences.

The two difficulties — that with respect to axioms and that with respect to primitives — arise from analogous asymmetries. In a system of Newtonian mechanics, given the initial conditions and masses of a system of particles, we can deduce univocally their paths. Given their paths, we may or may not be able to derive unique values for the masses. Given the laws and values of the generically defined primitives, we can deduce observation sentences; given any finite set of observation sentences, we

cannot generally deduce laws. When the matter is put in this way, the asymmetry is not surprising, and it is easy to see that the thesis of naive logical positivism — essentially the thesis of Condition (2) — is untenable unless it is weakened substantially.

Contextual Definitions, Implicit Definitions and Reduction Sentences.

Revisions of the concept of definition similar in aim to that discussed here have been proposed by a number of empiricists. Quine's [7, p. 42] notion of contextual definition, while nowhere spelled out formally, is an example:

The idea of defining a symbol in use was, as remarked, an advance over the impossible term-by-term empiricism of Locke and Hume. The statement, rather than the term, came with Frege to be recognized as the unit accountable to an empiricist critique. But what I am now urging is that even in taking the statement as unit we have drawn our grid too finely. The unit of empirical significance is the whole of science.

Braithwaite [1] carries the argument a step further by pointing out advantages of having in an empirical theory certain terms that are not uniquely determined by observations. His discussion of this point [1, pp. 76–77] is worth quoting:

We can, however, extend the sense of definition if we wish to do so. In explicit definition, which we have so far considered, the possibilities of interpreting a certain symbol occurring in a calculus are reduced to one possibility by the requirement that the symbol should be synonymous (within the calculus) with a symbol or combination of symbols which have already been given an interpretation. But the possibilities of interpreting a certain symbol occurring in a calculus may be reduced without being reduced to only one possibility by the interpretation already given of other symbols occurring in the formulae in the calculus. If we wish to stress the resemblance between the reduction of the possibilities of interpreting a symbol to only one possibility and the reduction of these possibilities but not to only one possibility, instead of wishing to stress (as we have so far stressed) the difference between these two sorts of reduction, we shall call the second reduction as well as the first by the name of definition, qualifying the noun by such words as "implicit" or "by postulate." With this extension of the meaning of definition the thesis of this chapter can be expressed by saying that, while the theoretical terms of a scientific theory are *implicitly defined* by their occurrence in initial formulae in a calculus in which there are derived formulae interpreted as empirical generalizations, the theoretical terms cannot be *explicitly defined* by means of the interpretations of the terms in these derived formulae without the theory thereby becoming incapable of growth.

As a final parallel, I will mention Carnap's concept of *reduction sentence* in his essay on *Testability and Meaning* [2, p. 442]. A reduction sentence for Q_3 is a sentence of the form, $Q_2 \supset (Q_1 \supset Q_3)$, where Q_2 is interpreted as the set of conditions under which the subsidiary implication holds, and where Q_1 is interpreted as a (partial) definiens for Q_3 . Thus, let Q_2 be the statement that a set of particles is isolated; Q_1 be the statement that a certain vector, \bar{m} , substituted for the coefficients in the equations stating the laws of conservation of momentum and angular momentum for the particles, satisfies those equations; and Q_3 be the statement that the components of \bar{m} are masses of the particles. Then $Q_2 \supset (Q_1 \supset Q_3)$ is essentially identical with the definition of mass in [8]. The subsidiary connective is an implication rather than an equivalence because there is no guarantee that another vector, \bar{m}' , may not also constitute a satisfactory set of masses, so that $Q_2 \supset (Q_1' \supset Q_3')$, where Q_1' is derived from Q_1 , and Q_3' from Q_3 by substituting \bar{m}' , for \bar{m} .

Definability Almost Everywhere. In preference to either definability or generic definability, we might want to have a term midway in strength between these two — a notion of definability that would guarantee that we could “usually” determine the defined term univocally, and that the cases in which we could not would be in some sense exceptional. Under certain conditions it is, in fact, possible to introduce such a term. *Suppose that B is a point in some space possessing a measure, and let there be a sentence of form (I) that holds almost everywhere in the space of B . Then, we say that a is DEFINED ALMOST EVERYWHERE.*

If, in [8], we take B as the time path of the system which satisfies the axioms in some interval $k < t < m$, and take the Lebesgue measure in the appropriate function space for the B 's as the measure function, then mass is defined almost everywhere, as is resultant force.

DEFINABILITY AND IDENTIFIABILITY

It has not generally been noted that the problem of definability of non-observables in axiomatizations of empirical theories is identical with what has been termed the “identification problem” in the literature of mathematical statistics [4, p. 70; 9, pp. 341–342]. The identification problem is the problem of estimating the parameters that appear in a system of equations from observations of the values of the variables in the same system of equations.

Some Types of Identifiability Problems. Consider, for example, a system of linear equations:

$$(1) \quad \sum_j a_{ij}x_j = b_i \quad (i = 1, \dots, n),$$

where the x 's are observables and the a 's and b 's are parameters. The a 's and b 's are generically defined by this system of equations, but they are not defined in Tarski's sense, for, no matter how many sets of observations of the x 's we have, the a 's and b 's are not uniquely determined. For suppose that A and b are a matrix and vector, respectively, that satisfy (1) for the observed x 's.³ Then A' and b' will also satisfy (1), where $A' = PA$ and $b' = Pb$ for any non-singular matrix P . To identify the a 's and b 's — that is, to make it possible to estimate them uniquely — additional constraints beyond those embodied in equations (1) must be introduced.

On the other hand, consider the system of linear difference equations:

$$(2) \quad \sum_j a_{ij}x_j(t) = x_i(t+1), \quad (i = 1, \dots, n)$$

where, as before, the x 's are observables, and the a 's and b 's constant parameters. In this case, the a 's are defined almost everywhere in the space of $\bar{x}(t)$. There are n^2 parameters to be estimated, and the number of equations of form (2) available for estimating them is $n(k-1)$, where k is the number of points in time at which the x 's are observed. Hence, for almost all paths of the system, and for $k > n+1$, the a 's will be determined uniquely.⁴

We see that the system of equations (2) is quite analogous to the system of equations used in [8] to define mass. In the latter system, for n particles, having $3n$ position coordinates, there are 6 second order differential equations (three for conservation of momentum, three for conservation of angular momentum) that are homogeneous in the m 's, and that must hold identically in t . There are $(n-1)$ parameters to be estimated — the number of mass-ratios of the particles, referred to a particular one of them as unit. Hence, for almost all paths of the system, the mass-ratios

³ In this entire discussion, we are disregarding errors of observation and the fact that the equations may be only approximately satisfied. For an analysis that takes into account these additional complications, the reader must refer to [3] and [4].

⁴ The convenience of replacing identifiability (equivalent to Tarski's definability) by almost-everywhere identifiability (equivalent to almost-everywhere definability) has already been noted in the literature on the identification problem [4, p. 82; 3, p. 53].

can be estimated uniquely from observations of the positions of the particles at $\left(\frac{n}{6} + 2\right)$ points in time.

Correspondingly, the system of equations (1) is analogous to the system of equations used in [8, p. 901] to define the component forces between pairs of particles. Component forces are only generically defined. Hence, although the masses of particles in a system and the resultant forces acting upon them can, in general, be estimated if there is a sufficient number of observations of the positions of the particles; the component forces cannot be so estimated unless additional identifying assumptions are introduced. Such additional assumptions might, for example, take the form of a particular force law, like the inverse square law of gravitational attraction.

Over-Identification and Testability. When a scientific theory is axiomatized with a view to clarifying the problems of testing the theory, a number of considerations are present that do not appear in axiomatizing deductive systems. Hence, it may be undesirable to imitate too closely the canons usually prescribed for the latter type of axiomatization. In addition to distinguishing primitive from defined terms, it may be advantageous to subdivide the former class as so to distinguish terms that are defined almost everywhere or that are only generically defined.

More fundamentally, whether particular terms are univocally determined by the system will depend not only on the specific sentences that have the form of definitions of these terms, but upon the whole set of sentences of the system. Our analysis of an actual axiom system for Newtonian particle mechanics bears out the contentions of Braithwaite and Quine that the definitions of non-observables often are, and must be, "implicit" or "contextual."

What does the analysis suggest, on the positive side, as a substitute for the too strict Condition (2)? In general, there will appear in an axiom system terms that are direct observables, and terms that are not. A minimum requirement from the standpoint of empiricism is that the system as a whole be over-identified: that there be possible sets of observations that would be inconsistent, collectively, with the sentences of the system. We have seen that this condition by no means guarantees that all the non-observables of the system will be defined terms, or even defined almost-everywhere.

A more radical empiricism would require that it be possible, by making

a sufficient number of observations, to determine uniquely the values of all parameters that appear in the system. To take a simple example, a strict interpretation of this condition would not permit masses to appear in the axiomatization of Newtonian mechanics, but only mass-ratios. Resultant forces would be admissible, but not component forces, unless sufficient postulates were added about the form of the force law to overdetermine them. We may borrow Quine's phrase for this requirement, and say that when it is satisfied for some set of terms, the terms are *defined contextually* by the system.⁵ The condition that all non-observables be defined contextually is still much weaker, of course, than the condition that they be defined.

For reasons of elegance, we may sometimes wish to stop a little short of insisting that all terms in a system be defined contextually. We have already mentioned a suitable example of this. In [8] mass ratios are defined almost everywhere, but masses are not defined contextually, even in an almost-everywhere sense. Still, we would probably prefer the symmetry of associating a mass number with each particle to a formulation that arbitrarily selected one of these masses as a numeraire.

Braithwaite has given us another reason, from the semantic side, for not insisting on contextual definition of all terms. He observes that if we leave some degrees of freedom in the system, this freedom allows us later to add additional axioms to the system, without introducing internal inconsistencies, when we have reason to do so. Thus, since the law of conservation of energy does not determine the zero of the temperature scale, the zero may be fixed subsequently by means of the gas laws.

Regardless of what position we take on empiricism in axiomatizing scientific theories, it would be desirable to provide for any axiom system theorems characterizing not only its syntactical properties (e.g., the independence, consistency, and completeness of the axioms), but its semantic properties (e.g., the degree of identifiability of its non-observables) as well.

⁵ Braithwaite's "implicit definition" will not do here, for he applies it specifically to the weaker condition of the previous paragraph.

Bibliography

- [1] BRAITHWAITE, R. B., *Scientific Explanation*. Cambridge 1955, XII + 376 pp.
- [2] CARNAP, R., *Testability and meaning*. Philosophy of Science, vol. 3 (1936), pp. 419–471, and vol. 4 (1937), pp. 1–40.
- [3] HOOD, W.. and T. C. KOOPMANS (eds.) *Studies in Econometric Method*. New York 1953, XIX + 323 pp.
- [4] KOOPMANS, T. C. (ed.), *Statistical Inference in Dynamic Economic Models*. New York 1950, XIV + 439 pp.
- [5] MCKINSEY, J. C. C., A. C. SUGAR and P. SUPPES, *Axiomatic foundations of classical particle mechanics*. Journal of Rational Mechanics and Analysis, vol. 2 (1953), pp. 253–272.
- [6] PADOA, A., *Essai d'une théorie algébrique des nombres entiers, précédé d'une introduction logique à une théorie déductive quelconque*. Bibliothèque du Congrès International de Philosophie, vol. 3 (1900).
- [7] QUINE, W., *From a Logical Point of View*. Cambridge (Mass.) 1953, VI + 184 pp.
- [8] SIMON, H. A., *The axioms of Newtonian mechanics*. Philosophical Magazine, ser. 7, vol. 33 (1947), pp. 888–905.
- [9] ———, *Discussion: the axiomatization of classical mechanics*. Philosophy of Science, vol. 21 (1954), pp. 340–343.
- [10] TARSKI, A., *Some methodological investigations on the definability of concepts*. Chapter 10 in *Logic, Semantics, Metamathematics*, Oxford 1956, XIV + 467 pp

AN AXIOMATIC THEORY OF FUNCTIONS AND FLUENTS

KARL MENDER

Illinois Institute of Technology, Chicago, Illinois, U.S.A.

The topic of this paper is a theory of some basic applications of mathematics to science. Part I deals with concepts of pure mathematics such as the logarithm, the second power, and the product, and with substitutions in the realm of those functions. Part II is devoted to scientific material such as time, gas pressure, coordinates — objects that Newton called *fluents*. Part III formulates articulate rules for the interrelation of fluents by functions. Properly relativized, the latter play that connective role for which Leibniz originated the term *function*.

I. FUNCTIONS

Explicitly, a real function with a real domain — briefly, a *function* — may be defined as a class of consistent ordered pairs of real numbers. Here and in the sequel, two ordered pairs of any kind are called *consistent* unless their first members are equal while their second members are unequal. If each pair $\in f_1$ (that is, belonging to the function f_1) is also $\in f_2$ — in symbols, if $f_1 \subseteq f_2$ — then f_1 is called a *restriction* of f_2 ; and f_2 an *extension* of f_1 . The *empty* function (including no pair) will be denoted by \emptyset . The class of the first (the second) members of all pairs $\in f$ is called the *domain* of f or $\text{dom } f$ (the *range* of f or $\text{ran } f$). If $\text{ran } f$ includes exactly one number, then f is said to be a *constant* function.

The following typographical convention ¹ will be strictly adhered to:

roman type for numbers; *italic type* for functions.

For instance, the logarithmic function — briefly, *log* — is the class of all pairs $(a, \log a)$ for any $a > 0$. The constant function consisting of all pairs $(x, 0)$ for any x will be denoted ² by O . The following are examples of a formula and a general statement, respectively: $\log e = 1$, and $\emptyset \subseteq f$ for any f . Here, $0, 1, e$ as

¹ Cf. Menger [10] referred to in the sequel as *Calculus*.

² Symbols for constant functions that are more elaborate than italicized numerals, such as c_1 and c_0 , must be used in order to express certain laws; e.g., that $c_{0+1} = c_0 + c_1$.

well as \log , O , and \emptyset are designations of specific entities, while a , x , and f are *variables* (i.e., symbols replaceable with the designations of specific entities according to the respective legends) — *number variables* or *function variables* as indicated typographically.

The intersection of any two functions is a function; e.g., that of \cos and \sin is the class of all pairs $((4n + 1)\pi/4, (-1)^n/\sqrt{2})$ for any integer n . The union of \cos and \sin , however is not a function. From the set-theoretical point of view, *functions do not constitute a Boolean algebra*³. But any two functions have a sum, a difference, a product, and a quotient provided $\frac{f_1}{f_2}$ is defined as the class of all pairs (x, q) such that $(x, p_1) \in f_1$, $(x, p_2) \in f_2$ and $\frac{p_1}{p_2} = q$ for some p_1 and p_2 — a definition that dispenses with any reference to zeros in the denominators. For instance,

$$\cot = \frac{1}{\tan}, \frac{f}{0} = \emptyset, \text{ and } \frac{f_1}{f_2} \cdot f_2 \subseteq f_1 \text{ for any } f, f_1, \text{ and } f_2.$$

Moreover, any function f_2 may be substituted into any function f_1 , the result $f_1 f_2$ (denoted by mere juxtaposition, whereas multiplication will always be denoted by a dot!) being the class of all pairs (x, z) such that $(x, y) \in f_2$ and $(y, z) \in f_1$ for some y . The *identity function*, i.e., the class of all pairs (x, x) for any x — an object of paramount importance — will be denoted⁴ by j . Its main property is bilateral neutrality under substitution:

$$(1) \quad jf = f = fj \text{ for any } f.$$

For each f , there is a *bilaterally inverse* function⁵, $\text{Inv } f$, which is the largest class of pairs (x, y) such that $(y, x) \in f$ and that, under substitution, $f \text{ Inv } f \subseteq j$ and $(\text{Inv } f)f \subseteq j$. For instance, $\text{Inv } j^3 = j^3$ and $\text{Inv } \exp = \log$. If j_+^2 is the class of all pairs (x, x^2) for any $x \geq 0$, then

$\text{Inv } j_+^2 = j^1$, $\text{Inv } j^1 = j_+^2$; similarly, $\text{Inv } j_-^2 = -j^1$, $\text{Inv } -j^1 = j_-^2$. But $\text{Inv } j^2$ consists of the single pair $(0, 0)$; and $\text{Inv } \cos = \emptyset$, while $\text{Inv } f$

³ For this reason, the postulational theories of binary relations, which are based on Boolean algebra (cf. especially, McKinsey [8] p. 85 and Tarski [22] p. 73), are inapplicable to functions.

⁴ Cf. *Calculus*. p. 74 and pp. 99–105. Cf. also Menger [11] and [12].

⁵ Cf. *Calculus*. pp. 91–95, where $\text{Inv } f$ is denoted by j/f . The fertility of this concept of inverse functions has been brought out by M. A. McKiernan's interesting and promising studies on operators. Cf. McKiernan [6], [7].

is a branch of *arccos* if $f \subseteq \cos$ and $\text{dom } f$ is the interval $[n\pi, (n+1)\pi]$, for some integer n .

In the traditional literature, the identity function has remained anonymous — one of the symptoms for the neglect of substitution in analysis. The usual reference—"the function x " — and the symbol x are complete failures in basic assertions. Even in order to assert substitutive neutrality, concisely expressed in (1), analysts are forced to introduce a better symbol than x — an ad hoc name of the identity function, say h — and then must resort to an awkward implication: If $h(x) = x$ for each x , then $h(f(x)) = f(x) = f(h(x))$ for any f and any $x \in \text{dom } f$.

The overemphasis on additive-multiplicative processes, which is characteristic of mathematics in the second quarter of this century, becomes particularly striking in passing from theories of functions based on explicit definitions to postulational theories — theories of rings of functions, of linear function spaces, etc., which stress those properties that functions or entities of any kind share with numbers. One of the few exceptions doing justice to substitution is the *trioperational algebra of analysis* ⁶. In it, functions are (undefined) elements subject to three (undefined) operations. With regard to the first two, denoted by $+$ and \cdot , the elements constitute a ring including neutral elements, 0 and 1 . The third, called substitution and denoted by juxtaposition, is associative and right-distributive with regard to the ring operations ⁷:

$$(2) \quad (f + g)h = fh + gh \text{ and } (f \cdot g)h = fh \cdot gh \text{ for any } f, g, h.$$

For many purposes, it is important to postulate a neutral element j satisfying (1).

Trioperational algebra has interesting applications to rings of polynomials as well as non-polynomials ⁸ but does not apply to the realm of *all* functions, even though the three operations can be defined for any two functions. The only ring postulate that is not generally satisfied is that, for each g , there exist an f such that $f + g = 0$. For instance, $-\log + \log$ is not 0 , but rather the restriction of 0 consisting of all $(x, 0)$ for $x > 0$. Only $f + \log \subseteq 0$ has solutions (namely, any $f \subseteq -\log$). What narrowed

⁶ Cf. Menger [13], [14].

⁷ In keeping with the traditional attitude toward substitution, the laws (2) are hardly ever mentioned even though they are as important in analysis as is the multiplicative-additive distributive law.

⁸ Cf. especially Milgram [18] p. 65, Heller [4] and Nöbauer [19].

the scope of trioperational algebra in its original form was the fact that it did not take the relation \subseteq into account.

A more satisfactory postulational approach to functions may be based on the following idea of a *hypergroup*: a set \mathcal{G} satisfying six postulates:

I. \mathcal{G} is partially ordered by a relation \subseteq . For some purposes it is convenient to assume that \mathcal{G} includes a (necessarily unique) *minimal* element \emptyset , such that $\emptyset \subseteq \gamma$ for each γ ; or, even further, that \mathcal{G} is *atomized* in the sense that (1) for each $\gamma \neq \emptyset$, at least one $\alpha \subseteq \gamma$ is an *atom* (i.e., such that $\alpha' \subset \alpha$ if and only if $\alpha' = \emptyset$); (2) $\gamma_1 \subseteq \gamma_2$ if and only if each atom $\subseteq \gamma_1$ is also $\subseteq \gamma_2$. For other purposes, \mathcal{G} may be assumed to be *intersectional*, i.e., to include, for any two elements γ_1 and γ_2 , a maximal element $\subseteq \gamma_1$ and $\subseteq \gamma_2$ — an *intersection*, $\gamma_1 \cap \gamma_2$.

II. In \mathcal{G} , there is an *associative operation*, \circ .

III. \mathcal{G} includes a *bilaterally and absolutely neutral element*, ν , such that $\gamma \circ \nu = \gamma = \nu \circ \gamma$ for any γ .

Clearly, ν is unique. The connection between \subseteq , \circ , and ν is established in the following postulate that simplifies the author's original development and is due to Prof. A. Sklar.

IV. $\gamma \subseteq \delta$ if there exists an element $\nu' \subseteq \nu$ such that $\nu' \circ \delta = \gamma$ and if and only if there exists an element $\nu'' \subseteq \nu$ such that $\gamma = \delta \circ \nu''$.

It readily follows that \circ is *bilaterally monotonic*; that is to say, $\gamma_1 \subseteq \gamma_2$ implies $\gamma_1 \circ \gamma \subseteq \gamma_2 \circ \gamma$ and $\gamma \circ \gamma_1 \subseteq \gamma \circ \gamma_2$ for any $\gamma_1, \gamma_2, \gamma$. If there is a minimal element, then \emptyset may be a *bilateral strict annihilator*:

$$\emptyset \circ \gamma = \emptyset = \gamma \circ \emptyset \text{ for each } \gamma.$$

Moreover, if $\nu_1 \subseteq \nu$ and $\nu_2 \subseteq \nu$, then $\nu_1 \circ \nu_2 \subseteq \nu_1$ and $\subseteq \nu_2$; thus, if \mathcal{G} is intersectional, $\nu_1 \circ \nu_2 \subseteq \nu_1 \cap \nu_2$.

V. For each γ , there exist two *unilaterally and relatively neutral elements*, $L\gamma$ and $R\gamma$ (the *left-neutral* and the *right-neutral* of γ) such that:

- 1) $L\gamma \circ \gamma = \gamma = \gamma \circ R\gamma$;
- 2) $L(\gamma_1 \circ \gamma_2) \subseteq L\gamma_1$ and $R(\gamma_1 \circ \gamma_2) \subseteq R\gamma_2$ for each γ_1 and γ_2 ;
- 3) if $\mu \subseteq \nu$, then $L\mu \subseteq \mu$ and $R\mu \subseteq \mu$.

Clearly, $L\gamma \subseteq \nu$ and $R\gamma \subseteq \nu$ for each γ . If $L\gamma = \gamma$ and/or $R\gamma = \gamma$, then $\gamma \subseteq \nu$. If $\gamma \subseteq \nu$, then $L\gamma = \gamma = R\gamma$. Hence $LL\gamma = RL\gamma = L\gamma$ for every γ . Moreover, $L\gamma = \emptyset$, $R\gamma = \emptyset$, and $\gamma = \emptyset$ are equivalent. If $\gamma \subseteq \delta$, then $L\gamma \subseteq L\delta$ and $R\gamma \subseteq R\delta$. If χ is an *annihilator* in the sense that $\gamma \circ \chi = L\chi$ and $\chi \circ \gamma = R\chi$ for each γ , and if $\chi \subseteq \nu$, then $\chi = \emptyset$. Moreover, $L\gamma$ and $R\gamma$ are characterized among the elements $\subseteq \nu$ by the following minimum property:

If $\mu \subseteq \nu$, then $\mu \circ \gamma = \gamma$ implies $L\gamma \subseteq \mu$, and $\gamma \circ \mu = \gamma$ implies $R\gamma \subseteq \mu$. It follows that $L\gamma$ and $R\gamma$ are unique for each γ . If \circ is commutative, then $L\gamma = R\gamma$ for each γ .

It will suffice, here, bypassing *unilaterally opposite* elements, to postulate finally

VI. For each γ , there is a *bilaterally opposite element* $\text{Op } \gamma$ such that $\text{Op } \gamma \circ \gamma \subseteq R\gamma$ and $\gamma \circ \text{Op } \gamma \subseteq L\gamma$ for each γ , and which, if one sets $\text{Op } \gamma \circ \gamma = R'\gamma$ and $\gamma \circ \text{Op } \gamma = L'\gamma$, has the following minimax property:

1) if $\delta \circ \gamma \subseteq R\gamma$ and $\gamma \circ \delta \subseteq L\gamma$, then $\delta \circ \gamma \subseteq R'\gamma$ and $\gamma \circ \delta \subseteq L'\gamma$;

2) if $\delta \circ \gamma = R'\gamma$ and $\gamma \circ \delta = L'\gamma$, then $\text{Op } \gamma \subseteq \delta$.

$\text{Op } \gamma$ is unique for each γ , and $L'\gamma \circ \gamma = \gamma \circ R'\gamma$, which might be called $C\gamma$, the *core* of γ . If $\mu \subseteq \nu$, then $\text{Op } \mu = R'\mu = L'\mu = C\mu = \mu$. For each atom, $\text{Op } \alpha$ is an atom, and $C\alpha = \alpha$. Additional assumptions would guarantee that

$\text{Op } \delta \circ \text{Op } \gamma \subseteq \text{Op}(\gamma \circ \delta)$; $\text{Op } \text{Op } \gamma \subseteq \gamma$; $\text{Op } \text{Op } \text{Op } \gamma = \text{Op } \gamma$ for each γ, δ . However, $\gamma \subseteq \delta$ does not imply $\text{Op } \gamma \subseteq \text{Op } \delta$.

An element γ of a hypergroup will be called *right-elementary* (or *left-elementary*) if $\delta \subset \gamma$ implies $R\delta \subset R\gamma$ (or $L\delta \subset L\gamma$). Each atom is bilaterally elementary — briefly, *elementary*. If \mathcal{G} is commutative and γ is unilaterally elementary, γ is elementary. If \mathcal{G} is atomized, then ϱ is right-elementary if and only if $\varrho \circ \mu = \varrho$ implies $\mu \subseteq \nu$. If, in contrast, $\kappa \circ \gamma \subseteq \kappa$ for each γ , then κ may be called a *left-annihilator*; and each κ' that is $\subseteq \kappa$, a *leftquasiannihilator*. Clearly, $\kappa' \circ \gamma$ and $\gamma \circ \kappa'$ are left-quasi-annihilators for any γ . If each element of \mathcal{G} is right-elementary (or elementary), then \mathcal{G} will be said to be *right-elementary*⁹ (or *elementary*).

With regard to addition as well as multiplication, the set of all functions is a commutative elementary hypergroup. The universal neutrals are 0 and 1 ; the relative neutrals of f are, as it were, vertical projections of f on 0 and 1 , respectively; the opposites of f are $-f$ and $\frac{1}{f}$.

With regard to substitution, the set of all functions is a (non-commutative) right-elementary hypergroup. The universal neutral is j . The relative neutrals, Rf and Lf , correspond to $\text{dom } f$ and $\text{ran } f$, respective-

⁹ Prof. B. Schweizer proposes to call δ a *right-neutralizer* of γ if $\gamma \circ \delta \subseteq \nu$ and $L\delta \subseteq R\gamma$, and to say that δ is (1) *maximal* if $\gamma \circ \delta' \subseteq \gamma \circ \delta$ for each right-neutralizer δ' (2) *saturating* if $\gamma \circ \delta = L\gamma$. One might then postulate that each element γ , on either side, has at least one maximal neutralizer or at least one saturating neutralizer.

ly.¹⁰ Thus the contrast between functions (classes of pairs of numbers) and their domains and ranges (classes of numbers) disappears. $\text{Op } f$ is $\text{Inv } f$. The left annihilators $\neq \emptyset$ are what may be called *universal constant functions*; the left-quasiannihilators $\neq \emptyset$ are the *constant functions*¹¹.

Another example of a hypergroup is the set of all binary relations in some universal set with regard to what logicians call the *relative product*¹². The universal neutral is the identity relation, while the relative neutrals again correspond to domains and ranges. $\text{Op } \gamma$ is a restriction — in general, a *proper* restriction — of the converse of the relation γ .

Geometrically, the situation may be interpreted in a set (a “*plane*” consisting of “*points*”) that is decomposed into mutually disjoint subsets (“*vertical lines*”). “*Simple*” sets, i.e., sets having at most one point in common with each vertical line, are the counterpart of functions. This vertical simplicity corresponds to right-side elementariness of functions. Substitution can be illustrated if, secondly, the plane is decomposed into disjoint subsets (“*horizontal lines*”) each of which has exactly one point in common with each vertical line; and if, thirdly, there is given a “*diagonal*” set having exactly one point in common with each vertical line as well as with each horizontal line. The diagonal corresponds to j ; each horizontal line, to a universal constant function; the vertical (the horizontal) projection of a simple set f on the diagonal, to Rf (to Lf); the points, to atoms. Fig. 1, p. 460, based on the assumption of ordinary vertical and horizontal lines and a straight diagonal, j , shows a simple plane construction¹³ of the result of substituting g into f . For any point a in the set g , move horizontally to j , then vertically to f , and finally horizontally back to the vertical line through a . The set of all points thus obtained is

¹⁰ In contrast to groups and hypergroups, a Brandt *groupoid* (Mathematische Annalen, vol. 96) only permits the composition of *some* elements. In “categories” (i.e., essentially, groupoids) of mappings of groups on groups, MacLane calls the one-side identities of a mapping its domain and range.

¹¹ In a self-explanatory way, one can say that functions constitute a commutative elementary *hyperfield* with regard to addition and multiplication (with the multiplicative annihilator 0) and (non-commutative) right-elementary hyperfields with regard to addition and substitution as well as multiplication and substitution. The functions may also be said to constitute a *trioperational hyperalgebra*.

¹² Cf., e.g., McKinsey [8] and Tarski [22].

¹³ Cf. *Calculus* pp. 89 ff. The traditional postulational theory of binary relations is inapplicable to functions (cf. ³). On the other hand, the plane construction of functional substitution here described may, as Prof. M. A. McKiernan observed, be utilized for binary relations instead of the 3-dimensional construction proposed by Tarski [22] pp. 78, 79.

f/g . In the figure, f has the shape of an exponential curve; g , that of $-j^2$; hence f/g , that of the probability curve.

Notwithstanding the analogy (brought out in the concept of a hypergroup) of addition and multiplication with substitution, the latter has a definite primacy. In an atomized non-commutative hypergroup \mathcal{G} , any binary operation \times (such as $+$ and \cdot), defined in the class of all atoms $\subseteq \nu$, may be extended to any two elements γ' and γ'' of \mathcal{G} by defining $\gamma' \times \gamma''$ as the minimum element including all atoms α such that there

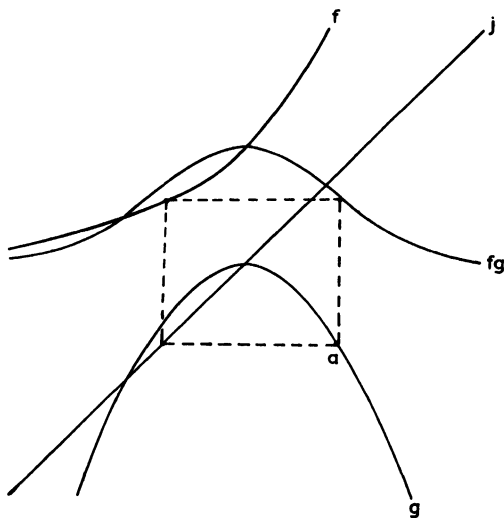


Fig. 1

exist two atoms $\alpha' \subseteq \gamma'$ and $\alpha \subseteq \gamma''$ satisfying the following conditions:

$$R\alpha = R\alpha' = R\alpha'' \text{ and } L\alpha = L\alpha' \times L\alpha''.$$

(This is essentially how the arithmetical operations are extended from numbers to functions.) Moreover, $\text{Neg } f$ and $\text{Rec } f$, the negative and the reciprocal of f , are obtainable by substituting f into $-j$ and j^{-1} , respectively; and, as will be shown presently, even $f + g$, $f \cdot g$, and $\frac{f_1}{f_2}$ can be obtained from 2-place functions S , P , and Q by substitution. In contrast,

¹⁴ In fact, no 2-place function yields fg even by substitution of f and g . Cf. *Calculus*, p. 304.

there are no functions of any kind which, for each f and g , would yield f/g or even $\text{Inv } f$ by additions and multiplications.¹⁴ Beyond any question, in the realm of functions substitution is the operation par excellence.

For any integer $m \geq 1$, a class of consistent pairs whose second members are real numbers while their first members are sequences of m real numbers is called, briefly, an *m-place function*¹⁵ and will be designated by a capital italic, except where lower case italics emphasize the 1-place character of functions such as those treated in what precedes. The class Q of the pairs $\left((x_1, x_2), \frac{x_1}{x_2} \right)$ for all x_1 and $x_2 \neq 0$ is a 2-place function. So are sum and product, S and P , from which, because of their associativity, an m -place sum and product, S^m and P^m , can be derived for each $m > 2$. Of particular importance is, for any two integers $1 \leq i \leq k$, the i -th k -place *selector function* I_i^k , that is, the class of all pairs $((x_1, \dots, x_k), x_i)$ for any x_1, \dots, x_k . Clearly, $I_1^1 = j$.

There are two main types of substitution of sequences of m functions into an m -place function (which, for $m = 1$, coincide with one another and with the substitution as defined on p. 455):

a) *product substitution*: $F^m[G_1, \dots, G_m]$, whose domain is a subset of the Cartesian product, $\text{dom } G_1 \times \dots \times \text{dom } G_m$, which is the class of all sequences $(\gamma_1, \dots, \gamma_m)$ for any $\gamma_1 \in \text{dom } G_1, \dots, \gamma_m \in \text{dom } G_m$.

b) *intersection substitution*: $F^m(G_1, \dots, G_m)$, whose domain is $\subseteq \text{dom } G_1 \cap \dots \cap \text{dom } G_m$. Unless G_1, \dots, G_m have the same place-number, that intersection is empty and $F^m(G_1, \dots, G_m) = \emptyset$; e.g., $P(j, S) = \emptyset$, while $P(I_1^2, S)$ and $P(I_2^2, S)$ are non-empty. Clearly,

$Q(f_1, f_2) = \frac{f_1}{f_2}$, as defined on p. 455; and $I_i^m(G_1, \dots, G_m) = G_i$ for any

m functions of the same placenumber. A simple generalization of the plane construction described on p. 459 to the 3-dimensional space¹⁶ yields $F^2(G_1^2, G_2^2)$.

Traditionally, $P(j^2, \log)$, $P[j^2, \log]$, $P(P, S)$, $P[P, S]$, $P(I_2^2, S)$, $P[j^2, S]$ are referred to as the functions $x^2 \log x$, $x^2 \log y$, $xy(x + y)$, $xy(u + v)$, $y(x + y)$, $x^2(y + z)$, respectively.

Either type of substitution can be extended to a realm of sequences

¹⁵ One might introduce numbers as 0-plane functions.

¹⁶ Cf. Menger [15] p. 224. Recently, S. Penner in his Master's thesis at Illinois Institute of Technology has extended the geometric axiomatics of substitution, outlined on p. 459 of the present paper, from 1-place to m -place functions in the $m + 1$ -dimensional space.

of functions. With each sequence, besides the number of functions in it, called the sequence-number, a placenumber will be associated. Either substitution of a second sequence into a first presupposes that *the sequence-number of the second be equal to the place-number of the first.*

a) An s-place *array* of r functions is a sequence such that the sum of the r places-numbers is s; for instance $\Phi_r^s = [F_1^{m_1}, \dots, F_r^{m_r}]$, where $s = m_1 + \dots + m_r$. Product substitution, defined by

$$\Phi_r^s \Psi_s^t = [F_1^{m_1}[G_1, \dots, G_{m_1}], \dots, F_r^{m_r}[G_{s-m_{r-1}+1}, \dots, G_s]],$$

clearly is associative and admits unilateral neutrals:

$$\Phi_r^s[\Psi_s^t X_t^u] = [\Phi_r^s \Psi_s^t] X_t^u \text{ and } |_r \Phi_r^s = \Phi_r^s = \Phi_r^s |_s^s,$$

where, for any $k \geq 1$, $|_k^k = [j]^k$, an array of k functions j.

b) An s-place *throw* of r functions is a sequence such that all r place numbers are s; for instance, $F_r^s = (F_1^s, \dots, F_r^s)$. Intersection substitution, defined by

$$F_r^s G_s^t = (F_1^s(G_1^t, \dots, G_s^t), \dots, F_r^s(G_1^t, \dots, G_s^t)),$$

is associative and admits unilateral neutrals. Let I_k^k be the k-place throw of all k-place selector functions in the natural order, and let $(I_k^k)^h$ denote the k-place throw of hk functions forming a chain of h throws I_k^k . Then

$$F_r^s(G_s^t H_t^u) = (F_r^s G_s^t) H_t^u \text{ and } I_r F_r^s = F_r^s = F_r^s I_s^s.$$

By means of intersection substitution, the array Φ_r^s and the throw F_r^s with the same components F_1^s, \dots, F_r^s can be connected: $F_r^s = \Phi_r^s(I_s^s)^r$.

Commutativity and associativity of addition and the distributive law can be expressed in the formulae:

$$S = S(I_2^2, I_1^2); S[j, S] = S[S, j]; P[S, j] = S[P, P](I_1^3, I_3^3, I_2^3, I_3^3).$$

The existence of right-neutrals has the following simple

Corollary. *Every non-empty function of any number of places lends itself to substitutions (of both types) with non-empty results.*

For any $k \geq 1$, the k-place throws of k functions form a hypergroup by intersection substitution. More generally, throws as well as arrays of functions constitute what might be called *hypergroupoids* — a concept that will be studied elsewhere.

Both types of substitution can be extended to n-ary relations. For instance, if P is a class of sequences of $n + 1$ elements; and if Π_1, \dots, Π_n are classes of (not necessarily consistent) ordered pairs, then $P[\Pi_1, \dots, \Pi_n]$ will denote the class of all sequences $(\alpha_1, \dots, \alpha_n, \gamma)$ such that for some β_1, \dots, β_n :

$$(\alpha_1, \beta_1) \in \Pi_1, \dots, (\alpha_n, \beta_n) \in \Pi_n \text{ and } (\beta_1, \dots, \beta_n, \gamma) \in P.$$

In what precedes, only *real* functions have been considered, but all statements (including the following remarks) remain valid if one selects a ring (or, where division is involved, a field) and writes *element of the ring (the field)* instead of *real number*.

The definitions of arithmetical operations for functions (addition, etc.) merely presuppose classes of consistent pairs *whose second members are real numbers*. The nature of the first members plays no role. Operating on functions with disjoint domains, however, yields \emptyset ; for instance, $j_{-2} + \log = \emptyset$ and $j.S = \emptyset$. Hence, for some results in a class of functions to be non-empty, it is necessary that *some domains be non-disjoint*¹⁷. With this proviso, the arithmetical operations may be extended to what I will call *functors*—classes of consistent quantities, if *quantity* is any ordered pair whose second member is a number¹⁸. Of course only functors whose domains consist of mathematical entities are objects of pure mathematics. Mathematical functors that are not functions have been called *functionals*; e.g., the class f_0^1 of all pairs $(f, f_0^1 f)$ for any integrable function f .

Substitution presents an altogether different situation. If the result $f_1 f_2$ is non-empty it is so because the first member of the pair $(y, z) \in f_1$ is the second member in a pair $(x, y) \in f_2$; in other words, because functions are classes of pairs *whose first and second members are of like nature*¹⁹. A similar reason accounts for substitutions of sequences of functions into functions of several places. In view of the corollary on p. 462, *the only functors that lend themselves to substitution with some non-empty results are the functions*. Calling every class of consistent quantities a “function” (which has been proposed) thus epitomizes overemphasis on addition and multiplication as well as supreme disregard for the paramount operation in the realm of functions — substitution.

II. FLUENTS

The objects of science and geometry to which Newton referred as fluents and which he and his successors have treated with supreme virtuosity

¹⁷ Functions of the same place-number, and even throws, satisfy this condition, and actually lend themselves to meaningful addition and multiplication.

¹⁸ Cf. *Calculus*, Chapter VII.

¹⁹ What that common nature of the elements *is* plays no role in the definition of substitution. For any set S , one may consider classes of consistent pairs of elements of S (self-mappings of S) and define substitution. Examples include n -place throws of n functions.

have not, in the classical literature, ever been defined either explicitly or, by postulates, implicitly. There are of course scientific procedures determining, for instance, $p\gamma_0$, the gas pressure in atm. of a specific instantaneous gas sample γ_0 , corresponding to arithmetical definitions of $\log 2$. But the function \log (even though its definition on p. 454 presupposes the understanding of $\log x$ for any x) must be distinguished from the numbers $\log x$ as well as from the class $\text{ran } \log$. Similarly, p — in the sequel, fluents as well as 1-place functions will be designated by lower case italics — must be distinguished from the numbers $p\gamma$ as well as from $\text{ran } p$ (the class of all those numbers). The fluent p is the class of all pairs $(\gamma, p\gamma)$ for any instantaneous gas sample γ .

Besides this (as it were, objective) pressure p , there is, for any observer A, a fluent p_A , *the gas pressure in atm. observed by A*, which is the class of all pairs $(\alpha, p_A\alpha)$ for any act α of A's reading a pressure gauge calibrated in atm., where $p_A\alpha$ denotes the number — *the pure number*, say, 1.5 — read by A as the result of α .

Thus extramathematical features (such as "denomination" and "dimension") that are often attributed to the values of p and p_A are, as it were, absorbed in the definitions of these fluents. Their values being pure numbers, also $\text{ran } p$ and $\text{ran } p_A$ are *objects of pure mathematics*. In contrast, $\text{dom } p$ and $\text{dom } p_A$ and, therefore, p and p_A themselves are *extramathematical objects*. The definition of an entire fluent adds to the knowledge of its values the idea of a class — a class that is highly significant in some physical laws and, in fact, indispensable if intuitive understanding (however efficient) of those laws is to crystallize in articulate formulations.

Differentiation between p , on the one hand, and the numbers $p\gamma$ or the class $\text{ran } p$, on the other, however slight the difference may appear, is at variance with the entire traditional literature on fluents inasmuch as the latter is at all articulate. McKinsey, Sugar, and Suppes²⁰ introduce time as a class of numbers (clock readings) and Artin²¹ takes a similar position (whereas, from the point of view here expounded, t_A , for an observer A, is the class of all pairs $(\tau, t_A\tau)$ for any act τ of clock reading performed by A). Courant says explicitly²² that Boyle's law deals only with the *values* of p and v and not with those quantities themselves. All that physics supplies, he emphasizes, are the classes of values of p and v .

In fact, Courant mentions p as an example of a *variable* (a symbol that

²⁰ Cf. McKinsey, Sugar and Suppes [9].

²¹ Cf. Artin [1], p. 70.

²² Cf. Courant [3], p. 16.

may be replaced with the designation of any element of a class of numbers), thereby illustrating another error pervading the traditional literature: the identification of fluents with what herein is called number variables, and the indiscriminate use of the term *variable* as well as the same (italic) type for both.

Yet — and this is a mere hint of the actual gulf separating the two — number variables *may* be interchanged, whereas fluents (e.g., abscissa and ordinate along a curve in a Cartesian plane, x being the class of all pairs $(\pi, x\pi)$ for any point π on the curve) *must not*. For instance, the class of all (x, y) such that $y = x^2$ is the same as the class of all (y, x) such that $x = y^2$, whereas the parabola $y = x^2$ and the parabola $x = y^2$ are different curves.

The confusion is enhanced by the use of the term variable, thirdly, for symbols that are replaceable with the designations of any element of some well-defined class of fluents or of classes of consistent quantities — in other words, for *fluent variables* or *c.c.q. variables*; e.g., for u in the statement $\frac{d \sin u}{du} = \cos u$ for any c.c.q. u that is continuous on (the limit class) $\text{dom } u$. Here, u may be replaced with the designation of the time²³ or the abscissa or even a continuous functional (as is f_0^1 in the realm of continuous functions whose limit is defined by uniform convergence), but nor with the designation of a number. One has

$$\frac{d \sin t}{dt} = \cos t, \quad \frac{d \sin x}{dx} = \cos x \quad \text{and even} \quad \frac{d \sin f_0^1}{df_0^1} = \cos f_0^1,$$

whereas $\frac{d \sin 1}{d1} = \cos 1$ is nonsense.

The literature also contains allusions to fluents that avoid confusing them with either classes of numbers or number variables. But those allusions (usually to “variable numbers”) are inarticulate beyond recognition. For instance Russell²⁴, Tarski²⁵, and other logicians in discussing number variables have repeatedly criticized the misconception of numbers that are capable of various values; and indeed, there are no numbers that are both 0 and 1, nor, as some one put it, numbers that have different values on weekdays and on Sundays. What logicians seem to overlook, however, is the fact that many obscure allusions to “variable numbers” do not refer to number variables in the logico-mathematical

²³ Strictly speaking, the domain of a fluent is not a limit class. In a model, however, according to the concluding remarks of the present paper, t and s may be assumed to be *continuous classes of consistent quantities on domains that are limit classes*. Cf. *Calculus*, pp. 220–225.

²⁴ Cf. Russell [20], p. 90.

²⁵ Cf. Tarski [21], pp. 3, 4.

sense, but rather represent utterly confused references to Newton's fluents. A fluent (without of course being a variable number) may indeed assume both the value 1 and the value 0. In fact, it may (as does, e.g., the admission fee in \$ to certain art galleries) assume the value 1 on weekdays and the value 0 on Sundays.

In the broadest sense, a *fluent* may be defined as a *class of consistent quantities with an extramathematical domain* — the c.'s c.q. with mathematical domains being functions and functionals. Fluents such as the class h of all pairs (F, hF) for any Frenchman F , where hF is F 's height in cm. (studied in biology and sociology), are sometimes called *variables*; their domains, *populations*.

Clearly, not every quantity, as defined on p. 463, is *interesting*; nor is every fluent *significant*, even if its elements are interesting quantities — think of the union of the height in the population of France and the weight in the population of Italy. Nor, for that matter, is every function and every functional important. While the general theory, of course, provides the scheme for handling *all* fluents, it is up to the individual investigator to apply it to some of the countless cases that are theoretically or practically significant.

Some critics of the theory here expounded have suggested that its basic idea, the concept of fluent, has always been known, viz., under the name of "real function" and, moreover, follows the pattern of Kolmogoroff's well-established concept of random variables — r.v.'s. Besides overextending the use of the term function (see p. 463), those critics seem to overlook: (1) that what is essential in the theory is the formulation of definitions for the (heretofore only intuitively used) concepts that Newton called fluents — definitions that are at variance with their traditional treatment, which ignores classes of pairs altogether (see p. 464); (2) that scientific fluents and r.v.'s lack one another's very characteristics and are, if anything, complementary rather than parallel concepts.²⁶ If A is a physical die, then the (*extramathematical*) class t_A of all pairs $(\delta, t\delta)$ for any act δ of rolling A is an *experimental fluent but not a r.v.* — not even if an additive functional ("probability") is defined for the 2^6 subsets of $\text{ran } t_A = \{1, \dots, 6\}$ (i.e., the class S of all possible outcomes of rolling A). On the other hand, in presence of such a probability functional on the subsets of S , any (*purely mathematical*) function having S as domain is a *r.v. but not a scientific fluent*; e.g., the function f for which $f(1) = \sqrt{7}$, $f(2) = \pi + e$, ..., $f(6) = \cos 2 + \log 5$. By their definitions, r.v.'s lack connections with experiment and observation. Again, scientific fluents such as t_A , gas pressure, and time lack the characteristic of r.v.'s, since the definition of a reasonable probability on subsets of their *domains* is completely out of the question. (What should be the probability of an act of rolling A , or of a gas sample or of an act of clock reading? Only in the *range* of a scientific fluent can one define frequency, relative frequency and, perhaps, probability.) (3) That even some-

²⁶ Cf. Menger [17], pp. 222–223.

one referring to all functors as "functions" cannot escape the use of a special term (say, "functions in the strict sense") referring to those functors whose domains consist of numbers or sequences of numbers. For (because of their substitutive properties, *not shared by any other functors*) these functors play a special role, and therefore are omnipresent in science as well as in mathematics. While in the light of the conceptual clarifications, terminological questions are quite insignificant, it does seem most appropriate to call *fluents* what Newton called fluents, and *functions*, what Leibniz called functions.

The union of non-identical fluents with the same domain is not a fluent. From a set-theoretical point of view, *fluents do not constitute a Boolean algebra*. One of the few positive formal properties of fluents is the possibility of substituting them into 1-place functions: $\log p$ is the class of all pairs $(\gamma, \log p\gamma)$ for any sample γ — a definition analogous to that of $\log \cos$. But while also the cosine of the logarithm is a c.c.q. $\neq \emptyset$, the pressure of the logarithm is empty. Every function permits *some* non-empty substitutions, whereas a fluent (like a functional) permits *none*.

Attempts have been made to dodge the problem of articulately connecting various fluents by *defining* some of them as functions of others²⁷. Yet, even if in a gallery a sign declares that admission costs \$ 1 on weekdays and is free on Sundays, the concept of admission fee cannot very well be said to be defined as a function of the time. Someone unfamiliar with that concept will not grasp it by reading the sign while, on the other hand, the concept is comprehensible to persons ignorant of the days of the week. Actually, admission fee might (for operative purposes) be defined as the class a of all pairs (A, aA) for any act of admitting a visitor, where aA is the amount in \$ charged during A . The sign, comprehensible only to those who know a and t , stipulates how the two are connected.

By substitutions into 2-place functions S , P , etc., significant addition, multiplication etc. of fluents can be defined, provided that their domains are non-disjoint — the only condition for arithmetical operations on c.'s c.q. to be non-empty (see p. 463). For instance, $P(p, v)$, the result of intersection substitution of p and v (whose common domain is the class of all γ) is $p.v$. But a slight change in the point of view raises difficulties. What, in view of the fact that $\text{dom } p_A$ and $\text{dom } v_A$ consists of acts of different (manometric and volumetric) observations, is the meaning of $p_A.v_A$? Since Boyle, it has become traditional to associate with that symbol (if only intuitively, i.e., without explicit definitions) the class $(\pi, \beta, p_A\pi.v_A\beta)$ for any two simultaneous acts π and β that A directs to

²⁷ Cf. the references in footnotes 20 and 21.

the same object; thus $p_A.v_A$ denotes the restriction of $P[p_A, v_A]$ to the class Γ of all pairs of *simultaneous and co-objective* acts $\in \text{dom } p_A \times \text{dom } v_A$. It thus appears that in operating on fluents, besides referring to the elements of their several domains, one may well have to relativize the operations to certain pairings of those domains. Such relativizations are imperative in formulating — articulately formulating — relations between fluents.

III. RELATIVE CONNECTIONS OF FLUENTS BY FUNCTIONS

Consider Boyle's law for gas undergoing an isothermal process — in proper units, $v = \frac{1}{p}$. If all that physics supplied were the values of v and p or the classes of those values, then Boyle might have discovered his law upon being presented with a bag containing cards each indicating a value of p , and another bag informing him of the values of v . But why, in that situation, should Boyle have paired each number in the first bag just with its reciprocal in the second rather than, say, with its square root? As a matter of fact, Boyle did not primarily pair numbers at all. Pairing numbers is what mathematicians do in defining functions. What Boyle actually paired were observations pertaining to the same object; and he discovered that

$$(3) \quad v\gamma = \frac{1}{p\gamma} \text{ for any inst. gas sample } \gamma \text{ at the fixed temperature.}$$

This statement is comparable to

$$(4) \quad \cot x = \frac{1}{\tan x} \text{ for any number } x \text{ that is not a multiple of } \pi/2,$$

with v and p corresponding to \cot and \tan ; and the sample variable γ , to the number variable x .

Unfortunately, the classical literature has done all that was possible to conceal the existing analogies. Besides, it has simulated a parallelism between v , p and x by indiscriminately referring to them as "variables" and using the same (italic) type for all of them (whereas the functions are usually denoted by \cot and \tan). In an attempt to mask the confusion between fluents and number variables, a contradiction in terms comparable to "enslaved freeman" was coined: "dependent variable". Finally, the true analogues of $v = 1/p$, formulae such as $\cot = 1/\tan$ (connecting two functions just as Boyle's law connects two

fluents) are anathema, and only the corresponding statements about numbers, such as (4) are admitted.²⁸

For an observer A, Boyle's law takes the form

$$(5) \quad v_{A\beta} = \frac{1}{p_{A\pi}} \text{ for any two acts } (\pi, \beta) \in \Gamma.$$

Relativizing connections of two fluents to a class Γ of pairs of simultaneous co-objective acts is very natural though not logically cogent. At any rate, since Galileo and Boyle, such (tacitly understood) relativizations have become second nature to physicists, who have transplanted them, as matters of course, even to quantum mechanics — a field where they are rather problematic. In $v = 1/p$, the pairing is altogether hidden.

On the level of general statements about fluents, however, the need for *explicit* relativizations is evident. The question "Is $w = 1/u$?" for any two fluents is incomplete. Certainly it does not necessarily refer to the entire class $\text{dom } u \times \text{dom } w$; that is to say, it does not necessarily mean "Is *each* value of w the reciprocal of *each* value of u ?" In this sense, for an affirmative answer it would be necessary that both u and w were constant fluents. The question thus must refer to some *subset* of $\text{dom } u \times \text{dom } w$. But to *which* subset? No particular subset of the Cartesian product of any two (especially disjoint) sets is or can be "natural". The intended subset must be *specified*. Such a relativization is necessary in order to make the question complete.

In the broadest sense, the connection of a class of consistent quantities w with another c.c.q. v relative to a set $\Pi \subseteq \text{dom } u \times \text{dom } w$ by the function f is described in the following basic definition:

$$w = fu(\text{rel. } \Pi) \text{ if and only if } (\alpha, \beta) \in \Pi \text{ implies } w\beta = fu\alpha.$$

Here, the consequent might be replaced with: $(u\alpha, w\beta) \in f$. For instance, (3) results if Π is the class I of all pairs (γ, γ) . The connection of *functions* by functions in traditional analysis is relative to restrictions of j . If j' is the restriction of j to numbers that are not multiples of $\pi/2$, then (4) subsumes under the general scheme:

$$\cot = j^{-1} \tan (\text{rel. } j') \text{ since } \cot y = j^{-1} \tan x \text{ for any } (x, y) \in j'.$$

²⁸ It is not unusual to write, e.g.: if $f = 1/g$, then $g = 1/f$ for any two functions f and g (thus dispensing with number variables). But, in violation of automatic substitutive procedures, the function variables f and g are replaced, e.g., by $\tan x$ and $\cot x$, and not in the traditional literature by \tan and \cot .

Clearly, $w = fu$ (rel. Π) implies $u = \text{Inv } fw$ (rel. conv. Π); and

$$w = fv \text{ (rel. } \Pi) \text{ and } v = gu \text{ (rel. } P) \text{ imply } w = fgu \text{ (rel. } \Pi P).$$

It is now clear why functions have been defined as on p. 454, and "multi-valued" functions have been strictly excluded. If the latter were admitted, then, relative to *every* pairing, *every* fluent would be a function of *every* other fluent. The question "Is w a function of u rel. Π ?", which is so important in science (e.g., in thermodynamics), would be deprived of any meaning. However, for any 2-place function F , one may define:

$$(6) \quad F(u, w) = 0 \text{ (rel. } \Pi) \text{ if and only if } (\alpha, \beta) \in \Pi \text{ implies } F(u\alpha, w\beta) = 0.$$

Of course only if $F(u\alpha, w\beta) \neq 0$ for *some* $(\alpha, \beta) \in \text{dom } u \times \text{dom } w$ (especially, if $F \neq 0$) does (6) establish a connection between u and w .

The most general connection of a functor w with n functors v_1, \dots, v_n relative to $P \subseteq \text{dom } v_1 \times \dots \times \text{dom } v_n \times \text{dom } w$ by the n -place function G is given by:

$$w = G[v_1, \dots, v_n] \text{ (rel. } P) \text{ if and only if}$$

$$(\beta_1, \dots, \beta_n, \gamma) \in P \text{ implies } w\gamma = G(v_1\beta_1, \dots, v_n\beta_n).$$

The chain rule reads as follows:

$$w = G[v_1, \dots, v_n] \text{ (rel. } P) \text{ and } v_i = F_i[u_{i,1}, \dots, u_{i,m}] \text{ (rel. } \Pi_i) \text{ imply}$$

$$w = G[F_1, \dots, F_n][u_{1,1}, \dots, u_{n,mn}] \text{ (rel. } P[\Pi_1, \dots, \Pi_n]).$$

The rate of change of w with, say, v_n rel. P (keeping v_1, \dots, v_{n-1} unchanged) is a fluent with the domain P , which must not be confused with the n -th place partial derivative $D_n G$, which is an n -place function with a domain $\subseteq \text{dom } G$. While the two symbols are frequently misrepresented as synonyms, the concepts are connected²⁹ by the formula:

$$\left(\frac{\partial w}{\partial v_n} \right)_{v_1, \dots, v_{n-1}} = D_n G(v_1, \dots, v_n) \text{ (rel. } P).$$

But it is important to note that the rate of change of w with v_n rel. P may well exist without w being a function of v_1, \dots, v_n rel. P . An analogous distinction is necessary between the cumulation of w with v_n and the n -th place partial integral of G .

From the preceding exposition of the material, based on explicit definitions, there emerge the outlines of its axiomatic treatment. A group

²⁹. Cf. *Calculus*, Chapter XI, especially pp. 306–315 and 332–341 and Menger [16].

I of postulates has to be devoted to *partial order* in a realm of undefined entities (called n -ary relations) in which there are two operations, *intersection* and *Cartesian multiplication*, subject to postulates of group II. In terms of these operations, associative substitutions are introduced (group III). Union of relations plays a small role, if any, and certainly none in that important subclass of relations whose elements are called *classes of consistent pairs* (group IV), because in the realm of c 's $c.p.$ union cannot in general be defined. Of particular significance among c 's $c.p.$ are *selector* and *identity relations* (group V) which, as has been illustrated in the realm of the 1-place functions, play the roles of domains of c 's $c.p.$ At this point, the class of all real numbers (or, if one pleases, a field or ring) enters the picture. By means of it, consistent classes of quantities or *functors* can be singled out (group VI) and, among them, *functions* constituting a hypergroupoid. Selector relations that are functions are the all-important selector functions, including the identity function j . What precedes is a basis for treating the connection of one functor with n other functors by means of an n -place function relative to an $n + 1$ -ary relation between their domains, as well as a functional interrelation of m functors relative to an m -ary relation.

Clearly, such an axiomatic theory represents the most general treatment of *models* in the sense in which this term is used in science, especially, in social sciences. An analogy appears with postulational geometry, which deals with undefined elements, called points and lines for the sake of a suggestive terminology, while all that is assumed about them is that they satisfy certain assumptions. Subsequently, they are compared with observable objects, e.g., in the astronomical space, with cross hairs and light rays. Models are formulated in terms of functor variables — undefined classes of consistent quantities, called, say, time and position or pressure and volume and denoted by t and s or p and v , for the sake of a suggestive terminology, while all that is assumed about them is that, relative to undefined pairings of their domains, those functors are interrelated by certain functions. Subsequently, an observer A compares them with observed fluents (t_A and s_A or p_A and v_A) relative to specified pairings of the domains of the latter. He trusts that, within certain limits of accuracy, the statements concerning the undefined functors in the model will be verified by known connections between the observed fluents — some of them, he hopes, by previously unknown connections³⁰.

³⁰ The ideas here outlined seem to supplement the existing theory on concept formation in empirical science; cf. Carnap [2] and Hempel [5].

As far as the general theory of fluents is concerned, the prediction may be ventured that indiscriminate uses of the term "variable" and of nondescript letters x will give way to more careful distinctions; and that references to domains of fluents as well as to pairings of those domains, once introduced, will be permanently incorporated in the articulate formulations of scientific laws.

Acknowledgements

The author is grateful to Professors M. A. McKiernan, B. Schweizer, and A. Sklar for valuable suggestions in connection with this paper, and to the Carnegie Corporation of New York for making it possible to devote time to the development of the material.

Bibliography

- [1] ARTIN, E., *Calculus and Analytic Geometry*. Charlottesville 1957, 126 pp.
- [2] CARNAP, R., *The methodological character of theoretical concepts*. In *Minnesota Studies in Philosophy of Science*, vol. 1, Minneapolis 1956.
- [3] COURANT, R., *Differential and Integral Calculus*, vol. 1,
- [4] HELLER, I., *On generalized polynomials*. Reports of a Mathematical Colloquium 2nd. ser., issue 8 (1947), pp. 58–60.
- [5] HEMPEL, C. G., *Fundamentals of concept formation in the empirical sciences*. International Encyclopedia of Unified Science, vol. 2 no. 7 Chicago 1952.
- [6] MCKIERNAN, M. A., *Les séries d'itérateurs et leurs applications aux équations fonctionnelles*. Comptes Rendus Paris, vol. 246 (1958), pp. 2331–2334.
- [7] —, *Le prolongement analytique des séries d'itérateurs*. Comptes Rendus Paris, vol. 246 (1958), pp. 2564–2567.
- [8] MCKINSEY, J. C. C., *Postulates for the calculus of binary relations*. Journal of Symbolic Logic, vol. 5 (1940), pp. 85–97.
- [9] —, SUGAR, A. C. and P. SUPPES, *Axiomatic Foundations of classical particle mechanics*. Journal of Rational Mechanics and Analysis, vol. 2 (1953), pp. 253–272.
- [10] MENDER, K., *Calculus. A Modern Approach*. Boston 1955, XVIII + 354 pp.
- [11] —, *The ideas of variable and function*. Proceedings of the National Academy, U.S.A., vol. 39 (1953) pp. 956–961.
- [12] —, *New approach to teaching intermediate mathematics*. Science, vol. 127 (1958) pp. 1320–1323.
- [13] —, *Algebra of Analysis*. Notre Dame Mathematical Lectures, vol. 3, 1944. 50 pp.
- [14] —, *Tri-operational algebra*. Reports of a Mathematical Colloquium, 2nd series, issue 5–6 (1945) pp. 3–10 and issue 7 (1946) pp. 46–60.
- [15] —, *Calculus. A Modern Approach*. Mimeographed Edition, Chicago 1952, XXV + 255 pp.

- [16] —, *Rates of change and derivatives*. Fundamenta Mathematicae, vol. 46 (1958), pp. 89–102.
- [17] —, *Random variables and the general theory of variables*. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, Berkeley 1956, pp. 215–229.
- [18] MILGRAM, A. N., *Saturated polynomials*. Reports of a Mathematical Colloquium, 2nd series, issue 7 (1946), pp. 65–67.
- [19] NÖBAUER, W., *Über die Operation des Einsetzens in Polynomringen*. Mathematische Annalen vol. 134 (1958) pp. 248–259.
- [20] RUSSELL, B., *The Principles of Mathematics*. Vol. 1. Cambridge 1903, XXIX + 534 pp.
- [21] TARSKI, A., *Introduction to Logic*. New York 1941, XVIII + 239 pp.
- [22] —, *On the calculus of relations*. Journal of Symbolic Logic, vol. 6 (1941) pp. 73–89.

AXIOMATICS AND THE DEVELOPMENT OF CREATIVE TALENT

R. L. WILDER

University of Michigan, Ann Arbor, Michigan, U.S.A.

Introduction. Perhaps I should apologize for presenting here a paper that embodies no new results of research in axiomatics. However, for some time I have felt that someone should record a description of an important method of teaching based on the axiomatic method, and this conference seems an appropriate place for it.

Actually, I can point to an excellent precedent in that the late E. H. Moore devoted most of his retiring address [2], as president of the American Mathematical Society, to a study of the role of the then rapidly developing abstract character of pure mathematics, especially the increasing use of axiomatics, in the teaching of mathematics in the primary and secondary schools. Just how much influence E. H. Moore's ideas had on the later developments in elementary mathematical education in this country, I do not know. It is perhaps significant that the increasing concern with these matters on the part of a large section of the membership of the American Mathematical Society (particularly in the Chicago Section) led, several years later, to the forming of a new organization, the Mathematical Association of America, whose special concern was with the teaching of mathematics in the undergraduate colleges.¹

Historical Development of the Method. We have heard a great deal, the past fifty years or so, of the use of the axiomatic method as a tool for research. Indeed, this use of the method has been justly considered as one of the most outstanding and surprising phenomena in the evolution of modern mathematics. Scarcely a half century ago, so great a mathematician as Poincaré could devote, in an article entitled *The Future of Mathematics* [6], less than half a page to the axiomatic method. And although conceding the brilliance of Hilbert's use of the method, he predicted that the problem of providing axiomatic foundations for various fields of mathematics would be very "restricted", and that "there would be nothing more to do when the inventory should be ended, which

¹ See [1], parts VII and XV 6, but especially p. 81 and p. 146.

could not take long. But when", he continued, "we shall have enumerated all, there will be many ways of classifying all; a good librarian always finds something to do, and each new classification will be instructive for the philosopher."

As recently as 1931, Hermann Weyl matched the contempt veiled in these remarks by a fear expressed as follows: "—I should not pass over in silence the fact that today the feeling among mathematicians is beginning to spread that the fertility of these abstracting methods [as embodied in axiomatics] is approaching exhaustion. The case is this: that all these nice general notions do not fall into our laps by themselves. But definite concrete problems were conquered in their undivided complexity, single-handed by brute force, so to speak. Only afterwards the axiomaticians came along and stated: Instead of breaking in the door with all your might and bruising your hands, you should have constructed such and such a key of skill, and by it you would have been able to open the door quite smoothly. But they can construct the key only because they are able, after the breaking in was successful, to study the lock from within and without. Before you can generalize, formalize and axiomatize, there must be a mathematical substance. I think that the mathematical substance in the formalizing of which we have trained ourselves during the last decades, becomes gradually exhausted. And so I foresee that the generation now rising will have a hard time in mathematics."²

Evidently mathematical genius does not correlate well with the gift of prophecy, since neither Poincaré's disdain nor Weyl's fears have been justified. Neither of these eminent gentlemen seems to have realized that a powerful creative tool was being developed in the new uses of the axiomatic method. It was Weyl's good fortune to live to see and acknowledge the triumphs of the method. And undoubtedly had Poincaré lived to observe how the method contributed to the progress of mathematics, he would gladly have admitted his prophetic shortcomings. It is easy to comprehend why they felt as they did, and as, conceivably, a majority of their colleagues felt. For until quite recent years, the method had achieved its most notable successes in geometry, where axiom systems often served as suitable embalming devices in which to wrap up theories already worked out and in a stage of decline. The value of the method as a tool for opening up vast new domains for mathematical

² Quoted from H. Weyl [7]. It is to Weyl's credit that he acknowledges, in this connection, the brilliant results obtained by Emmy Noether by her pioneering use of the axiomatic method in algebra.

investigation, as it has done in algebra and topology for example, was not yet sufficiently exemplified to make an impression on the mathematical public. Peano's fundamental researches in logic and number theory were concealed in his unique "pasigraphy"; and besides, was not this again a case of wrapping old facts in new dress (mused the uncomprehending analyst)? Similarly Grassmann's earlier work in his (now justly appreciated) *Ausdehnungslehre* was concealed in a mass of philosophical obscurities, and moreover the philosophy of the time was dominated by a Kantian intuitionism not receptive to the idea of mathematics as a science of formal structures.

Nevertheless, the evolution of modern mathematics was proceeding in a direction which made inevitable those uses of axiomatics with which every modern mathematician is now familiar. Noone, among the mathematicians active around the turn of the century, appears to have been more aware of this trend than the American mathematician E. H. Moore. Moore's interest in, and use of, axiomatic procedures is well known, and I have already remarked on his interest in the influence which they might have on the teaching of elementary mathematics. Of importance for my purposes is the influence of his ideas on a group of young mathematicians who were under his tutelage at the time, particularly R. L. Moore and O. Veblen. Both Veblen and R. L. Moore wrote their doctoral dissertations in the axiomatic foundations of geometry. And the interests of both soon turned to what was at the time a new branch of geometry in which metric ideas play no official role, viz. *topology*, or as it was then called, *analysis situs*.

It is an interesting fact, however, that the topological interests of the two diverged, the one, Veblen, following the line initiated by Poincaré and subsequently called "combinatorial topology", the other, R. L. Moore, following the line stemming from the work of Cantor and Schoenflies and subsequently called "set-theoretic topology". And whereas the latter, set-theoretic topology, lent itself naturally to the axiomatic approach which Moore continued to develop, the former, combinatorial topology, was not left by Poincaré (whose feelings toward the axiomatic method we have already indicated above) in a form suitable to axiomatic development.

The first major work of R. L. Moore in "analysis situs" [3], was published in 1916.³ It embodied a set of axioms characterizing the analysis

³ There are three axiom systems given in this paper. In our remarks we refer only to that one which is designated in [3] by the symbol " Σ_1 ".

situs of the euclidean plane. In a later paper [4], Moore showed this axiom system to be categorical, and still later [5] applied it in a way prophetic of the new, creative uses of the axiomatic method soon to come into vogue.

However, of much more importance for my present purposes, was the manner in which Moore ⁴ used his axiom system for plane analysis situs for discovering and developing creative talent. Those of us who are accustomed to the use of axioms in constructing new theories, or for other technical creative purposes, may have lost sight of the fact that the axiomatic method can serve as the basis for a most useful *teaching device*.

I am not referring to the traditional use of axioms in teaching high school geometry of the euclidean type. Although here, in the hands of an inspired teacher, the method can and sometimes undoubtedly does turn up potential mathematicians, most of the teaching of high school geometry seems to be of two kinds. Either it is based on the use of a standard text book in which the theorems are all worked out in detail for study by the pupil, with a supply of minor problems — so-called “originals” — to be done by the pupil and geared usually to the ability of the “average” student; or it is carried out in connection with a laboratory process which is supposed to exemplify the so-called “reality” of the theorems proved, thereby preventing the abstract character of the system from becoming too dominant. In short, the whole process may be considered overly adapted to the capacities of the “average” student and consequently generally loses — perhaps justifiably — its potentiality for developing the mathematical talents of the more gifted student.

Nor am I referring to the fact that quite commonly, in our graduate courses in algebra and topology, we use the axiomatic method for setting up abstract systems. I mean something *more* than this. What I mean can perhaps be indicated by a remark which one of my former students made to me in a recent letter: “I am having quite good success teaching a course, called Foundations of Analysis, by the Moore-Socrates method.” The use by Moore of the axioms for plane analysis situs in his teaching had many elements in common with the Socratic method as revealed in the “Dialogues”, especially in the general type of interplay between master and pupil.

Moore proceeded thusly: He set up a course which he called “Foundations of Mathematics”, and admitted to attendance in the course only

⁴ From here on, by “Moore” I shall mean R. L. Moore.

such students as he considered mature enough and sufficiently sympathetic with the aims of the course to profit thereby. It was not, then, a required course, nor was it open to any and all students who wanted to "learn something about" Moore's work. He based his selection of students, from those applying for admission, on either previous contacts (usually in prior courses) or (in the case of students newly arrived on the campus) on analysis via personal interview — usually the former (that is, previous contacts). The amazing success of the course was no doubt in some measure due to this selection process.

He started the course with an informal lecture in which he supplied some explanation of the role to be played by the undefined terms and axioms. But he gave very little intuitive material — in fact only meagre indication of what "point" and "region" (the undefined terms) might refer to in the possible interpretations of the axioms. He might take a piece of paper, tear off a small section, and remark "Maybe that's a region". However, as the course progressed, more intuitive material was introduced, oftentimes by means of figures or designs set up by the students themselves.

The axioms were eight ⁵ in number, but of these he gave only two or three to start with; enough to prove the first few theorems. The remaining axioms would be introduced as their need became evident. He also stated, without proof, the first few theorems, and asked the class to prepare proofs of them for the next session.

In the second meeting of the class the fun usually began. A proof of Theorem 1 would be called for by asking for volunteers. If a valid proof was given, another proof different from the first might be offered. In any case, the chances were favorable that in the course of demonstrating one of the theorems that had been assigned, someone would use faulty logic or appeal to a hastily built-up intuition that was not substantiated by the axioms.

I shall not bore you with all the details; you can use your imaginations, if you will, regarding the subsequent course of events. Suffice it to say that the course continued to run in this way, with Moore supplying theorems (and further axioms as needed) and the class supplying proofs. I could give you many interesting — and amusing — accounts of the byplay between teacher and students, as well as between the students themselves; good-natured "heckling" was encouraged. However, the point to be emphasized is that Moore *put the students entirely on their own*

⁵ One of these was later shown (by the present author [8]) not to be independent.

resources so far as supplying proofs was concerned. Moreover, there was no attempt to cater to the capacities of the "average" student; rather was the pace set by the *most talented* in the class.

Now I grant that there seems to be nothing sensational about this. Surely others have independently initiated some such scheme of teaching.⁶ The noteworthy fact about Moore's work is that he began finding the capacity for mathematical creativeness where no one suspected it existed! In short, he *found* and *developed creative talent*. I think there is no question but that this was in large measure due to the fact that the student felt that he was being "let in" on the management and handling of the material. He was afforded a chance to experience the thrill of creating mathematical concepts and to glimpse the inherent beauty of mathematics, without having any of the rigor omitted in order to ease the process. And in their turn, when they went forth to become teachers, these students later used a similar scheme. True, they met with varying success — after all, a pedagogical system, no matter how well conceived, must be operated by a good teacher. Their success was striking enough, however, that one began to hear comments and queries about the "Moore method". And it is partly in response to these that I am talking about the subject today. It seems that it is time someone described the method as it really operated, and perhaps thereby cleared up some of the folklore concerning it.

Description of the Method. In the interest of clarity, I shall arrange my remarks with reference to certain items which I think, after analyzing the method, are in considerable measure basic to its success. These items are as follows:

1. *Selection of students capable* (as much as one can tell from personal contacts or history) *of coping with the type of material to be studied.*
2. *Control of the size of the group participating; from four to eight students probably the best number.*
3. *Injection of the proper amount of intuitive material, as an aid in the construction of proofs.*
4. *Insistence on rigorous proof, by the students themselves, in accordance with the ideal type of axiomatic development.*
5. *Encouragement of a good-natured competition; it can happen that as many different proofs of a theorem will be given as there are students in the class.*

⁶ Professor A. Tarski informed me after the reading of this paper that he had used a somewhat analogous method in one of his courses at the University of Warsaw.

6. *Emphasis on method, not on subject matter.* The amount of subject matter covered varies with the size of class and the quality of the individual students.

I think these six items lie at the heart of the method. Of course they slight the details; e.g., the manner in which Moore exploited the competition between students, and the way in which he would encourage a student who seemed to have the germ of an idea, or put to silence one who loudly proclaimed the possession of an idea which upon examination proved vacuous. I imagine that it was in such things as these that Moore most resembled Socrates. But these are matters closely related to Moore's personality and capability as a teacher, so I shall confine myself to the six points enumerated above so far as the description of the method is concerned. They are, I realize, themselves pedagogic in nature, but more of the nature of what I might call *axiomatic pedagogy*. They constitute, I believe, a guide to the successful use of axiomatics in the development of creative talent.

I would like to comment further on them:

1. *Selection of students capable of coping with the type of material to be studied.* I have already made some remarks in this regard. I pointed out that Moore based his judgment regarding maturity either on his experience with the student in prior courses, or on personal interviews. I might add, parenthetically, that as the years went by and his students began to use his methods in their own teaching, a sort of code developed between them whereby one of the "cognoscenti" would apprise one of his colleagues in another university of the availability of potentially creative material. For example, the "pons asinorum" of Moore's original axiom system was "Theorem 15". If one of Moore's graduates wished to place a student for further work under the tutelage of another of Moore's students at a different institution, and could include in his recommendation the statement, "He proved Theorem 15", then this became a virtual "open sesame".

But Moore, himself, was not dependent on other institutions; he found his students, generally speaking, in the student body of the University of Texas. He had a singular ability for detecting talent among undergraduates, and often set his sights on a man long before he was ready for graduate work. Indeed, in some instances, he would allow in his class in "Foundations" an undergraduate whom he deemed ready for creative work. For Moore believed that a man should start his creative work as

soon as possible, and the younger the better. He reasoned that one could always pick up "breadth" as he progressed. It was not unusual for him to discover talent in his calculus classes. And once he suspected a man of having a potentially mathematical mind, he marked that man for the rest of the course as one with whom he would cross his foils, so to speak. By the end of the term, he was usually pretty sure of his opinion of the man.

Of course he could not, in the very nature of the case, always be certain. This applies especially to those who entered his class as graduate students from other institutions, who had had no previous work with him, and whom he had to screen usually in a single interview at registration time. And when a student of little or no talent did slip by, he was doomed to a semester of either sitting and listening (usually with little comprehension), or to feverishly taking notes which he hoped to be able to understand by reading outside of class. In the latter case he was often disappointed, for as we all know, one's first proof of a theorem is usually not elegant, to understate the case, and the first proofs of a theorem given in class were likely to be of this kind. But as I stated before, the aim of the course was not so much to give certain *material* — the student who wished the latter would have been better advised to read a book or to seek out the original material in journals. I would call these "note-takers" the "casualties" of the course. So you see it was *humane*, as well as good strategy, to allow only the "fit" to enroll in the course.

I might remark, too, that those of us who went from Texas to other institutions as young instructors, did not usually find it possible to institute Moore's "exclusion policy" in all its rigor. For various reasons, we often had to throw our courses open to one and all. This naturally led to certain modifications, as, for instance, making sure that the "note-takers" ultimately secured an elegant proof; this seemed the least that they were entitled to under a system where they were not sufficiently forewarned of what to expect, and of especial importance if the material covered was to be used by the student as basic information in later courses.

2. *Control of the size of the group participating; from four to eight probably the best number.* This is obviously not independent of 1., since Moore's method of selecting students was clearly suited to keeping down the size of the class. Some of us, however, especially during periods of high enrollments, have had to cope with classes of as many as 30 students or more. I can report from experience that even with a class this large, the

method can be used. Of course inevitably a few (sometimes only two or three) students "star in the production". I have found, however, that these "star" students often profited from having such a large audience as was afforded by the "non-active" portion of the class. Often the "non-stars" came up with some good questions and sometimes — rarely to be sure — with a suggestion that led to startling consequences.

In short, although from four to eight is the ideal size of class for the use of the axiomatic method, it is not impossible to handle classes of as many as 30 while using the method.

3. *Injection of the proper amount of intuitive material, as an aid in the construction of proofs.* This, I hardly need emphasize, must be handled carefully. With no intuitive background at all, the student has little upon which to fix his imagination. The undefined terms and the axioms become truly meaningless, and a mental block perhaps ensues. Here the instructor must exercise real ingenuity, striving to furnish that amount of intuitive sense that will be sufficient to suggest processes of proof, while at the same time holding the student to the axiomatic basis as a foundation for all assertions of the proof.

I have always been interested, in my use of the axiomatic method in Topology, in observing the degree to which the various students used figures in giving a demonstration. Some relied heavily on figures; others used none at all, being content to set down the successive formulae of the proof. I have noticed that the former type of student usually developed an interest in the geometric aspects of the subject, following the tradition of classical topology, while the latter developed greater interest in the new algebraic aspects of the subject. There may be considerable truth in the old folklore that some are naturally geometric-minded, while others have not so much geometric sense but show great facility for algebraic types of thinking. I don't know of any better way to discover a student's propensities in these regards, than to give him a course in modern topology on axiomatic lines.

4. *Insistence on rigorous proof, by the students themselves, in accordance with the ideal type of axiomatic development.* I want to emphasize here two advantages that the axiomatic development offers.

In the first place, I have seen the method rescue potentially creative mathematicians from oblivion. Without knowing the reason therefor, they had become discouraged and depressed, having taken course after

course without "catching on" — with no spark of enlightenment. The reason for this was evidently that their innate desire for clearcut understanding and rigor was continually starved in course after course. One can appreciate the gleam in a student's eye when, provided with the type of rigor which the axiomatic method affords, he finds his mathematical self at last; for the first time, seemingly, he can let his creative powers soar with a feeling of security. This is truly one of the ways in which creative talent is discovered.

In the second place, even the average student feels happy about knowing just what he is allowed to assume, and in the feeling that at last what he is doing has, in his eyes, an almost perfect degree of validity. I can illustrate by an example here. I was once giving a course in the structure of the real number system, using a system of axioms and the "Moore method". In the class was a man who had virtually completed his graduate work and was writing his dissertation in the field of analytic functions. At the end of the course he came to me and said, "You know, I feel now for the first time in my life, that I really *understand* the theory of real functions". I knew what had happened to him. Despite all his courses and reading in function theory, he had never felt quite at ease in the domain of real numbers. Now he felt that, having been thrown wholly on his own resources, he had come to grips with the most fundamental properties of the real number system, and could, so to speak, "look a set of real numbers in the eye!"

5. *Encouragement of a good-natured competition.* I have found that an interesting by-play often developed between students, either to see who could first obtain the proof of a theorem, or failing that, who could give the most elegant proof. I presume this is a foretaste of the situation in which the seasoned mathematician often finds himself. I hardly need to cite historic instances to an audience like this; instances in which a settlement of a long outstanding problem was clearly in the offing, and the experts were vying with one another to see who would be the first to achieve the solution. This always adds zest to the game of mathematics, either on the elementary level or on the professional level. And no system of teaching lends itself better to this sort of thing than the one I am discussing.

There is also the possibility that an original-minded student will discover a new and more elegant proof of a classical theorem. I have had this happen several times, and on at least one occasion, to which I shall

refer again below, the proof given failed to use one of the conditions stated in the classical hypothesis, so that a new and stronger theorem resulted.

6. *Emphasis on method, not on subject matter.* When one lectures, or uses a text, the student is frequently presented with a theorem and then given its proof before he has had time to digest the full meaning of the theorem. And by the time he has struggled through the proof presented, he has been utterly prejudiced in favor of the methods used. They are all that will occur to him, as a rule. Use of the axiomatic method with the student providing his own proof, forces an acquaintance with the *meaning* of the theorem, and a *decision* on a method of proof. I have continually in my classes, whenever existence proofs were demanded, urged the students to find constructive methods whenever possible. In this way, I have had presented to me constructive proofs in instances where I did not theretofore know that such proofs could be given.

In short, use of the axiomatic method not only encourages the student to develop his own creative powers, but sometimes leads to the invention of new methods not previously conceived.

There is one other feature of the method as Moore used it that I have omitted above, for various reasons, chiefly because of the vagueness of its terms and the debatability of any interpretation of it:

7. *Selection of material best suited to the method.* It is probably wisest to select certain special subjects which seem best suited for the avowed purposes of discovering and developing creative ability. For example, one might select material that presupposes little in the way of special techniques (as, for instance, the techniques of classical analysis), but that does require that ability to think abstractly which should be a characteristic of the mature student of mathematics, and which requires little intuitive background. The material which Moore chose was of this nature; another such selection might be the theory of the linear continuum.

In the case of the material which Moore selected, the student was led quickly to the frontiers of knowledge; that is, to the point where he might soon be doing original research. I think this aspect of his method is not, however, essential to its success in developing creative talent. As Moore used the method, the line between what was known and what was unknown was not revealed to the students. Customarily they were not apprised of the source of the axioms or the theorems; for all they knew, these had probably never been published. And he could go on with them

to unsolved problems through the device of continuing to state theorems whose validity he might not himself have settled, without their ever being aware of the fact.

Consequently, so far as this item 7 is concerned, I would say that the important aspect of it is the selection of material requiring little intuitive background and presupposing mathematical maturity but little technique. The techniques of deduction, proof, and of discovering new theorems are naturally part of the design of the course; the axiomatic method is ideal for the development of these, and they should be given priority over the quantity of material covered.

The justification for the system is of course its success. It soon reveals to both teacher and student whether or not the latter possesses mathematical talent. It quickly selects those who have the "gift", so to speak, and develops their creative powers in a way that no other method ever succeeded in doing. Every mathematician, now and in the past, has recognized the necessity for doing mathematics, not just reading it, and has assigned "originals" for the student to do on his own. In the Moore system, we find the "original" par excellence — there is nothing in the course but originals! I should repeat, in connection with these remarks, that it is not unusual for a student to find a new proof of a known theorem that deserves publication, as well as for new theorems to be found. I had one outstanding case of this in my own use of the method, where the new proof showed one could dispense with part of the traditional hypothesis; and I had the student go on to incorporate his methods into proving another and similar theorem which was historically related to the former and was susceptible to the same improvement in its hypothesis.

The fact that what the logician would call the "naive" axiomatic method is used, does not seem to cause any objection from the student. In fact, I am afraid that a strict formalism might not work so well; although this is debatable, and certainly a carefully formulated proof theory would be quite adaptable to certain types of material. The use of a "natural" language throughout, except for the technical undefined terms, was, however, an important feature of the method as Moore used it, not only aiding the intuition but enabling that competition mentioned in item 5 to "wax hot" at crucial moments.

This brings me to some remarks about an area of teaching in which tradition is most strong, namely the *undergraduate curriculum*. Today we hear a great deal about encouraging the young student to go into a

mathematical or scientific career. Unfortunately much potential creative talent is lost to mathematics early in the undergraduate training, and much of this, I am sure, is due to traditional modes of presentation. It is possible that the axiomatic approach offers at least a partial solution of this problem.

The axiomatic method in the undergraduate course. As Moore used the axiomatic method for teaching on the graduate level, the aim was to discover and develop creative ability. Is there not a possibility that the method could be employed to advantage at a lower level, so that the potentially creative mathematician will be encouraged to continue in mathematics to the point where his talents can be more decisively put to the test?

I am convinced that one of our greatest errors in the United States educational system has been to underestimate the ability of the young student to *think abstractly*. Moreover, I am convinced that as a result, we actually force him to think "realistically" where actually he would *prefer* to think abstractly, so that by the time he begins graduate work, his ability to abstract has been so dulled that we have to try to develop it anew.

It seems probable that we could try using the axiomatic method on a lower level, perhaps even on the freshman level, at selected points where the material is of a suitable kind. In the interests of caution, perhaps we should experiment on picked groups first, as well as with carefully selected material. It is possible that we might light creative sparks where, with the conventional type of teaching, no light would ever dawn. Some years ago I had a chance to do this sort of thing, with a picked group of around ten students. I did not have an opportunity to teach most of these students again until they became graduates. But I am happy to state that a majority of them went on to the doctorate — not necessarily in mathematics, for some turned to physics — but at least they went on into creative work. I don't wish to give myself credit here; it is the *method* that deserves the credit. These men discovered unsuspected powers in themselves, and could not resist cultivating and exercising them further. Moreover, I found they were delighted at being able to establish their ideas on a rigorous basis. For example, in starting the calculus, I gave them precise definitions, etc., for a foundation of the theory of limits in the real number system, and let them establish rigorously on this foundation all the properties of limits needed in the calculus. The result was

that they covered the calculus in about half the time ordinarily required. Admittedly some of this saving in time was due to the select nature of the class, but a major part, I am convinced, was due to the confidence and interest induced by establishing the theory of limits on a firm basis.

In the presidential address of E. H. Moore to which I referred in my introduction, he stressed the advisability of mixing the real and the abstract in the teaching of mathematics in the secondary schools. But (and here I quote from E. H. Moore's address, p. 416) "— when it comes to the beginning of the more formal deductive geometry why should not the students be directed each for himself to set forth a body of geometric fundamental principles, on which he would proceed to erect his geometrical edifice? This method would be thoroughly practical and at the same time thoroughly scientific. The various students would have different systems of axioms, and the discussion thus arising naturally would make clearer in the minds of all precisely what are the functions of the axioms in the theory of geometry." Here was evidently a suggestion for the creative use of axiomatics at the high school level.

There are currently experiments being conducted in some undergraduate colleges which are based on modifications of the methods Moore used. For example, I know of one case ⁷ where a special course of this kind, for freshmen, has been devised. One-half the course is spent establishing arithmetic, on an axiomatic basis. The numbers 0, 1, 2, etc. are used, but the development is rigorous, and indeed approaches the rigor of a formal system in that the *rules for proof* are explicitly set forth. By the use of variables, the student is led gradually into algebra, which occupies most of the latter half of the course. The course terminates in an analysis, based on truth tables, of the formal logic to which the student has gradually become accustomed during the course. I judge that one of the reasons for the success which the course seems to have achieved is that the student is made aware of the reasons for the various arithmetic manipulations in which he was disciplined in the elementary schools; as, for instance, why one inverts and multiplies in order to divide by a fraction. This course has, incidentally, revealed that students who do not do well on their placement examinations are not necessarily laggards, weak-minded, or susceptible of any of the other easy explanations, but that they often are intelligent, capable persons who have been antagonized by traditional drill methods. Moreover, some of these students are induced by the course into going further in mathematics. I believe this course is

⁷ At the University of Miami.

still in a developmental stage, and I await with interest reports on its effectiveness. One gets the feeling from reading the text used that the student is being treated with trust, as naturally curious to know the *why* of what he is doing, and as being intelligent enough to find out if permitted!

During the past few years there has been published a number of elementary texts which use the axiomatic method to some extent. Perhaps this is a sign of a trend. I hope that in my remarks I have not over-emphasized to such an extent as to give an impression that I think the axiomatic method is a *cure-all*. I do not think so. Nor do I think it desirable that all courses should be axiomatized! But I believe that the great advances that the method has made in mathematical research during the past 50 years can, to a considerable extent, find a parallel in the teaching of mathematics, and that its wise and strategic use, at special times along the line from elementary teaching to the first contacts with the frontiers of mathematics, will result in the discovery and development of much creative talent that is now lost to mathematics.

Bibliography

- [1] ARCHIBALD, R. C., *A semicentennial history of the American Mathematical Society 1888–1938*. American Mathematical Society Semicentennial Publications, vol. 1, New York 1938, V + 262 pp.
- [2] MOORE, E. H., *On the foundations of mathematics*. Bulletin of the American Mathematical Society, vol. 9 (1902–03), pp. 402–424.
- [3] MOORE, R. L., *On the foundations of plane analysis situs*. Transactions of the American Mathematical Society, vol. 17 (1916), pp. 131–164.
- [4] ———, *Concerning a set of postulates for plane analysis situs*. Transactions of the American Mathematical Society, vol. 20 (1919), pp. 169–178.
- [5] ———, *Concerning upper semi-continuous collections of continua*. Transactions of the American Mathematical Society, vol. 27 (1925), pp. 416–428.
- [6] POINCARÉ, H., *The foundations of science*. Lancaster, Pa., 1946, XI + 553 pp.
- [7] WEYL, H., *Emmy Noether*. Scripta Mathematica, vol. 3 (1935), pp. 1–20.
- [8] WILDER, R. L., *Concerning R. L. Moore's axioms Σ_1 for plane analysis situs*. Bulletin of the American Mathematical Society, vol. 34 (1928), pp. 752–760.

